SIAM J. MATH. ANAL.
Vol. 33, No. 1, pp. 1–15

SELF-SIMILAR LATTICE TILINGS AND SUBDIVISION SCHEMES[*]

DING-XUAN ZHOU[†]

Dedicated to Professor Gil Strang on the occasion of his 65th birthday

Abstract. Let $M \in \mathbb{Z}^{s\times s}$ be a dilation matrix and let $\mathcal{D} \subset \mathbb{Z}^s$ be a complete set of representatives of distinct cosets of $\mathbb{Z}^s/M\mathbb{Z}^s$. The self-similar tiling associated with M and $\mathcal{D}$ is the subset of $\mathbb{R}^s$ given by $T(M,\mathcal{D}) = \{\sum_{j=1}^{\infty} M^{-j}\alpha_j : \alpha_j \in \mathcal{D}\}$. The purpose of this paper is to characterize self-similar lattice tilings, i.e., tilings $T(M,\mathcal{D})$ which have Lebesgue measure one. In particular, it is shown that $T(M,\mathcal{D})$ is a lattice tiling if and only if there is no nonempty finite set $\Lambda \subset \mathbb{Z}^s \setminus (\mathcal{D}-\mathcal{D})$ such that $M^{-1}((\mathcal{D}-\mathcal{D})+\Lambda)\cap\mathbb{Z}^s \subset \Lambda$. This set Λ can be restricted to be contained in a finite set K depending only on M and $\mathcal{D}$. We also give a new proof for the fact that $T(M,\mathcal{D})$ is a lattice tiling if and only if $\cup_{n=1}^{\infty}(\sum_{j=0}^{n-1} M^j(\mathcal{D}-\mathcal{D})) = \mathbb{Z}^s$. Two approaches are provided, one based on scrambling matrices and the other based on primitive matrices. These will follow from the characterization of subdivision schemes associated with nonnegative masks in terms of finite powers of finite matrices, without computing eigenvalues or spectral radii. Our characterization shows that the convergence of the subdivision scheme with a nonnegative mask depends only on the location of its positive coefficients.

Key words. self-similar lattice tilings, subdivision schemes, column-stochastic matrices, scrambling matrices, primitive matrices

AMS subject classifications. 41A15, 42C40, 15A51

PII. S0036141000367977

1. Introduction.

Self-similar tilings are defined in terms of dilation matrices and digit sets. A dilation matrix M in $\mathbb{R}^s (s \in \mathbb{N})$ is an $s \times s$ integer matrix with all the eigenvalues greater than 1 in modulus, i.e., $\lim_{n\to\infty} M^{-n} = 0$. A digit set $\mathcal{D}$ associated with the dilation matrix M is a complete set of representatives of distinct cosets of the quotient group $\mathbb{Z}^s/M\mathbb{Z}^s$. If $m = |\det M|$, then $\mathcal{D}$ consists of m elements $\{\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{m-1}\}$ and $\varepsilon_i - \varepsilon_j \notin M\mathbb{Z}^s$ for $i \neq j$.

The *self-similar tiling* associated with a dilation matrix M and a digit set $\mathcal{D}$ is defined to be the subset $T(M,\mathcal{D})$ of $\mathbb{R}^s$ as

$$T(M,\mathcal{D}) := \left\{ \sum_{j=1}^{\infty} M^{-j}\alpha_j : \quad \alpha_j \in \mathcal{D} \right\}. \tag{1.1}$$

Self-similar tilings have been studied in a variety of contexts in the literature; see, e.g., [6, 5, 10, 11, 20] and the references therein.

The measure of a self-similar tiling is always a positive integer. Here we are interested in the case when this measure $\text{meas}(T(M,\mathcal{D}))$ is exactly one. If this happens, we call $T(M,\mathcal{D})$ a *self-similar lattice tiling* because the integer translates of $T(M,\mathcal{D})$ tile the space $\mathbb{R}^s$ without overlapping. These lattice tilings provide useful examples not only for fractal geometry but also for wavelet analysis. How to characterize self-similar lattice tilings in terms of the digit sets is our main concern here.

[*] Received by the editors February 17, 2000; accepted for publication (in revised form) December 5, 2000; published electronically April 12, 2001. This research was supported by the Research Grants Council of Hong Kong.

http://www.siam.org/journals/sima/33-1/36797.html

[†] Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong (mazhou@math.cityu.edu.hk).

In the univariate case, $M = m > 1$ is a positive integer and $\mathcal{D}$ consists of m integers $\{\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{m-1}\}$ satisfying $\varepsilon_j \equiv j (\mathrm{mod} m)$ for $j = 0, 1, \ldots, m-1$. Then $T(M, \mathcal{D})$ has measure one if and only if the numbers $\varepsilon_1 - \varepsilon_0, \ldots, \varepsilon_{m-1} - \varepsilon_0$ are relatively prime, i.e., $\gcd(\varepsilon_1 - \varepsilon_0, \ldots, \varepsilon_{m-1} - \varepsilon_0) = 1$. This result was proved independently by Gröchenig and Haas [5] and by Zhou [21].

In the multivariate case, things are more difficult. Gröchenig and Madych [6] used the Cohen condition for orthogonality of refinable functions and obtained many interesting examples of self-similar lattice tilings. A useful necessary condition for self-similar lattice tilings was provided by Lagarias and Wang [10]: if $T(M, \mathcal{D})$ has measure one, then the smallest M-invariant sublattice $\mathbb{Z}[M, \mathcal{D}]$ of $\mathbb{Z}^s$ containing the difference set $\mathcal{D} - \mathcal{D} := \{\alpha - \beta : \alpha, \beta \in \mathcal{D}\}$ is $\mathbb{Z}^s$. A necessary and sufficient condition was given by Gröchenig and Haas [5] in terms of the spectrum of a finite matrix. To apply this condition, one needs to check whether some eigenvalues of the matrix have modulus 1 exactly, which is somehow numerically instable. By the tiling theorem in [11], $T(M, \mathcal{D})$ is a self-similar lattice tiling if and only if $\cup_{n=1}^{\infty} \sum_{j=0}^{n-1} M^j (\mathcal{D} - \mathcal{D}) = \mathbb{Z}^s$. A new proof for this fact will be given in this paper by means of subdivision schemes.

In this paper we show that $T(M, \mathcal{D})$ is a lattice tiling if and only if there is no nonempty finite set $\Lambda \subset \mathbb{Z}^s \setminus (\mathcal{D} - \mathcal{D})$ such that $M^{-1}\big((\mathcal{D} - \mathcal{D}) + \Lambda\big) \cap \mathbb{Z}^s \subset \Lambda$. This set Λ can be restricted to be contained in a finite set K depending only on M and $\mathcal{D}$. The lattice tiling property is also equivalent to $K \subset \sum_{j=0}^{n-1} M^j (\mathcal{D} - \mathcal{D})$ for some $1 \leq n \leq 1 + \#\big(K \setminus (\mathcal{D} - \mathcal{D})\big)$; see Theorem 5. We also provide some criteria for checking self-similar lattice tilings in terms of finite powers of finite matrices. Two approaches will be presented, one based on scrambling matrices (section 3) and the other based on primitive matrices (section 4). Our criteria will be consequences of general characterizations of L_2-convergence of subdivision schemes associated with nonnegative masks. Observe that the characteristic function of the self-similar tiling (1.1) satisfies the refinement equation

$$\phi(x) = \sum_{\alpha \in \mathcal{D}} \phi(Mx - \alpha), \qquad x \in \mathbb{R}^s.$$

The symbol of the corresponding mask is a conjugate quadrature filter. Hence the self-similar lattice tiling can be studied by the convergence of subdivision scheme associated with this equation.

2. Subdivision schemes and transfer operators.

2. Subdivision schemes and transfer operators. Subdivision schemes are often used to solve *refinement equations* of the form

$$\phi(x) = \sum_{\alpha \in \mathbb{Z}^s} a(\alpha) \phi(Mx - \alpha), \qquad x \in \mathbb{R}^s, \tag{2.1}$$

where $a := \{a(\alpha)\}_{\alpha \in \mathbb{Z}^s}$ is a finitely supported sequence called the *refinement mask*. We restrict a to be real-valued in this paper. When $\sum a(\alpha) = m$, the refinement equation (2.1) has a unique compactly supported distributional solution under the normalized condition $\hat{\phi}(0) = 1$ (e.g., [1, 2]). Here $\hat{\phi}$ denotes the Fourier transform of ϕ and ϕ is called the *normalized solution* of (2.1).

In order to find a solution in L_p, we start with the initial function $\phi_0(x) = \Pi_{j=1}^s \varphi(x_j)$, where $x = (x_1, \ldots, x_s)$ and φ is the univariate hat function supported and given on $[0, 2]$ by $\varphi(t) = \min\{t, 2 - t\}$. Then the subdivision scheme associated with (2.1) is defined to be a sequence of functions $\{\phi_n\}$ as

$$\phi_n(x) = \sum_{\alpha \in \mathbb{Z}^s} a(\alpha) \phi_{n-1}(Mx - \alpha), \qquad x \in \mathbb{R}^s, \quad n \in \mathbb{N}.$$

We say that the *subdivision scheme* associated with (2.1) converges in $L_p (1 \leq p \leq \infty)$ if and only if $\{\phi_n\}$ converges in L_p:

$$\lim_{n\to\infty} \|\phi_n - \phi\|_p = \lim_{n\to\infty} \left\{ \int_{\mathbb{R}^s} |\phi_n(x) - \phi(x)|^p dx \right\}^{1/p} = 0.$$

To characterize the L_p-convergence of the subdivision scheme in terms of the mask, one needs the concept of the p-norm joint spectral radius; see, e.g., [8, 2]. It is hard to compute the p-norm joint spectral radius. However, when p is an even integer, Zhou showed in [22] (see also [19]) that the p-norm joint spectral radius of a finite set of matrices can be computed exactly and explicitly by the spectral radius of a single finite matrix.

It is well known (see, e.g., [4, 8]) that the L_2-convergence of subdivision schemes can be characterized in terms of the spectral radius of a finite matrix derived from the transfer operator which was introduced to wavelet analysis by Lawton [12]. In the multivariate case, such a characterization was given independently by Han and Jia [7], Lawton, Lee, and Shen [13], and Strang [18]. To state this fact, let b be the sequence defined by

$$b(\alpha) = \frac{1}{m} \sum_{\beta\in\mathbb{Z}^s} a(\beta)a(\alpha+\beta), \qquad \alpha \in \mathbb{Z}^s.$$

Obviously, the sequence b is finitely supported on $\operatorname{supp} a - \operatorname{supp} a$. The transfer operator is associated with the bi-infinite matrix $(b(M\alpha-\beta))_{\alpha,\beta\in\mathbb{Z}^s}$. To get the finite matrix, we restrict both α, β to be in a finite index set $K \subset \mathbb{Z}^s$ such that the space of all sequences supported in K, $\ell(K)$ is invariant under the operator $(b(M\alpha - \beta))$. Also, we require that K contains Ω, *the support of the sequence* b. Such a set K exists. For example, by [7] we may take

$$K = \mathbb{Z}^s \cap \left(\sum_{n=1}^{\infty} M^{-n}(\Omega \cup M\Omega) \right).$$

Now we have a finite set K, and the sequence b satisfies

$$b(M\alpha - \beta) \neq 0, \quad \beta \in K \qquad \Longrightarrow \qquad \alpha \in K.$$

The finite matrix we need for the L_2-convergence is

$$F := \big(b(M\alpha-\beta)\big)_{\alpha,\beta\in K}. \tag{2.2}$$

A necessary condition for the convergence of the subdivision scheme is the sum rule of order one: $\sum_{\alpha\in\mathbb{Z}^s} a(M\alpha+\beta) = 1$ for every $\beta \in \mathbb{Z}^s$. Under this condition,

$$\sum_{\alpha\in\mathbb{Z}^s} b(M\alpha+\beta) = \frac{1}{m} \sum_{\gamma\in\mathbb{Z}^s} a(\gamma) \sum_{\alpha\in\mathbb{Z}^s} a(M\alpha+\beta+\gamma) = 1 \qquad \forall \beta \in \mathbb{Z}^s.$$

If a sequence v supported on K ($v \in \ell(K)$) satisfies $\sum v(\alpha) = 0$, then the vanishing of $b(M\alpha - \beta)$ for $\beta \in K, \alpha \notin K$ yields

$$\sum_{\alpha\in K} (Fv)_\alpha = \sum_{\alpha\in K}\sum_{\beta\in K} b(M\alpha-\beta)v(\beta) = \sum_{\beta\in K} \left\{ \sum_{\alpha\in\mathbb{Z}^s} b(M\alpha-\beta) \right\} v(\beta) = 0.$$

That means the space

$$V := \left\{ v \in \ell(K) : \sum_{\alpha \in K} v(\alpha) = 0 \right\} \tag{2.3}$$

is invariant under the action of F.

With the above notation, we can now state the following known result [7, 13, 18] on the L_2-convergence of subdivision schemes.

THEOREM A. *Let M be a dilation matrix and let $a := \{a(\alpha)\}_{\alpha \in \mathbb{Z}^s}$ be a finitely supported sequence with $\sum a(\alpha) = m$. Then the subdivision scheme associated with* (2.1) *converges in L_2 if and only if*

(a) $\sum_{\alpha \in \mathbb{Z}^s} a(M\alpha + \beta) = 1 \; \forall \beta \in \mathbb{Z}^s$*;*

(b) $\rho(F|_V) < 1$*, i.e., the spectral radius of the finite matrix F restricted to the invariant subspace V is less than* 1.

This is a very nice characterization. However, one still has to compute the eigenvalues of the matrix, which is not so stable numerically. In particular, for our purpose of self-similar tilings, the mask will be nonnegative. In this case, F is a column-stochastic matrix, i.e., F is nonnegative and for every $\beta \in K$

$$\sum_{\alpha \in K} F_{\alpha,\beta} = \sum_{\alpha \in \mathbb{Z}^s} b(M\alpha - \beta) = 1.$$

Thus, F would have eigenvalues other than 1 on the unit circle, if condition (b) of Theorem A is not true. This difficulty can be overcome by using the special property caused by the nonnegative mask.

3. The L_2-convergence of subdivision schemes with nonnegative masks.

In this section we characterize the L_2-convergence of the subdivision scheme associated with a nonnegative mask in terms of finite powers of the column-stochastic matrix F. This approach depends mainly on a result of *scrambling* column-stochastic matrices used by Jia and Zhou in [9].

A matrix $A = (A_{\alpha,\beta})_{\alpha,\beta \in K}$ is called *column-stochastic* if all its entries are nonnegative and

$$\sum_{\alpha \in K} A_{\alpha,\beta} = 1 \qquad \forall \beta \in K.$$

We say that the column-stochastic matrix A is *scrambling* if each pair of columns has positive entries in some common row. It is well known that AB is column-stochastic if both A and B are, and AB is scrambling if B is scrambling and A is column-stochastic. Denote $\|v\|_1$ as the ℓ_1-norm of sequences. The result from [9] that we need here is the following.

LEMMA. *Let $A = (A_{\alpha,\beta})_{\alpha,\beta \in K}$ be column-stochastic and let V be defined by* (2.3)*. Then*

$$\|A|_V\|_1 := \sup_{0 \neq v \in V} \frac{\|Av\|_1}{\|v\|_1} < 1$$

if and only if A is scrambling.

We are now in a position to state the main result of this section.

THEOREM 1. *Let M be a dilation matrix and let $a := \{a(\alpha)\}_{\alpha \in \mathbb{Z}^s}$ be a finitely supported* nonnegative *sequence with $\sum a(\alpha) = m$. Set K and F as in Theorem A.*

Let $N := \#K$. Then the subdivision scheme associated with (2.1) *converges in L_2 if and only if*

(i) $\sum_{\alpha\in\mathbb{Z}^s} a(M\alpha+\beta) = 1\ \forall\beta\in\mathbb{Z}^s$*;*

(ii) *there exists some integer n with $1 \leq n \leq (3^N - 2^{N+1}+1)/2$ such that the matrix F^n is scrambling.*

Proof. The sufficiency follows from Theorem A, our lemma, and the well-known fact that

$$\rho(F|_V) = \lim_{k\to\infty} \|F^k|_V\|_1^{1/k} = \inf_{k\in\mathbb{N}} \|F^k|_V\|_1^{1/k}.$$

To see the necessity, we suppose that the subdivision scheme converges in L_2. Then condition (i) holds by Theorem A. Theorem A also tells us that

$$\rho(F|_V) = \inf_{k\in\mathbb{N}} \|F^k|_V\|_1^{1/k} < 1.$$

Hence, $\|F^k|_V\|_1 < 1$ for some $k \in \mathbb{N}$. Since F^k is column-stochastic, by our lemma, F^k is scrambling. Observe that condition (ii) is equivalent to the fact that $F^{(3^N-2^{N+1}+1)/2}$ is scrambling. It is sufficient for us to show that $F^{(3^N-2^{N+1}+1)/2}$ is scrambling.

Suppose to the contrary that the column-stochastic matrix $F^{(3^N-2^{N+1}+1)/2}$ is not scrambling. Then there are two distinct elements γ_1 and γ_2 in K such that for every $\alpha\in K$ either $(F^{(3^N-2^{N+1}+1)/2})_{\alpha,\gamma_1}$ or $(F^{(3^N-2^{N+1}+1)/2})_{\alpha,\gamma_2}$ is zero. Let $0 \leq j \leq (3^N - 2^{N+1}+1)/2, \gamma\in\{\gamma_1,\gamma_2\}$. We have

$$(F^{(3^N-2^{N+1}+1)/2})_{\alpha,\gamma} = \sum_{\beta\in K} (F^{(3^N-2^{N+1}+1)/2-j})_{\alpha,\beta}(F^j)_{\beta,\gamma}.$$

Therefore, for every $\beta\in K$, either $(F^j)_{\beta,\gamma_1}$ or $(F^j)_{\beta,\gamma_2}$ is zero, since for some $\alpha\in K$ depending on β, $(F^{(3^N-2^{N+1}+1)/2-j})_{\alpha,\beta} \neq 0$. Here $F^0 = I$. Define

$$I_{j,\gamma} := \{\beta\in K : (F^j)_{\beta,\gamma} > 0\}.$$

Then we know that $I_{j,\gamma} \neq \emptyset$ and for $0 \leq j \leq (3^N-2^{N+1}+1)/2, I_{j,\gamma_1}\cap I_{j,\gamma_2} = \emptyset$. The number of different unordered pairs of disjoint nonempty subsets of K is $(3^N - 2^{N+1} + 1)/2$ (see [16]). Hence, there must exist some $1 \leq p < l \leq (3^N - 2^{N+1}+1)/2$ such that $I_{p,\gamma_1} = I_{l,\gamma_1}$ and $I_{p,\gamma_2} = I_{l,\gamma_2}$. That means, for every $\beta\in K, \gamma\in\{\gamma_1,\gamma_2\}, (F^p)_{\beta,\gamma}$ and $(F^l)_{\beta,\gamma}$ are either both positive or both zero.

Let $q\in\mathbb{N}, \alpha\in K, \gamma\in\{\gamma_1,\gamma_2\}$. Then

$$(F^{p+q(l-p)})_{\alpha,\gamma} = \sum_{\beta\in K} (F^{(q-1)(l-p)})_{\alpha,\beta}(F^l)_{\beta,\gamma}$$

vanishes if and only if

$$(F^{p+(q-1)(l-p)})_{\alpha,\gamma} = \sum_{\beta\in K} (F^{(q-1)(l-p)})_{\alpha,\beta}(F^p)_{\beta,\gamma}$$

equals zero. Hence, for $\gamma\in\{\gamma_1,\gamma_2\}, q\in\mathbb{N}, I_{p+q(l-p),\gamma} = I_{p,\gamma}$. It follows that for any $q\in\mathbb{N}$,

$$I_{p+q(l-p),\gamma_1} \cap I_{p+q(l-p),\gamma_2} = I_{p,\gamma_1}\cap I_{p,\gamma_2} = \emptyset.$$

Thus the column-stochastic matrix $F^{p+q(l-p)}$ is not scrambling. Choose q such that $p+q(l-p) > k$. Then we conclude that F^k is not scrambling, which is a contradiction. This tells us that condition (ii) holds. □

In [15, 3, 9] the uniform convergence of subdivision schemes associated with nonnegative masks was characterized in terms of finite products of m matrices. This in connection with the autocorrelation of the mask yields another way to characterize the L_2-convergence. However, we need a set of m matrices and many more matrix products need to be checked.

The argument for the bound $(3^N - 2^{N+1} + 1)/2$ was provided by Paz [16], who proved that $(3^N - 2^{N+1} + 1)/2$ is a sharp bound for the power in checking scrambling products of several matrices. Here we give a complete proof for the reader's convenience. For a single matrix, this bound can most likely be largely reduced. With the approach given in the next section (Theorem 2) we will need to check only the powers up to $(N-1)^2 + 1$ for our purpose of checking the L_2-convergence. However, Theorem 1 plays an important role in deriving a nice characterization of self-similar lattice tilings (Theorem 5) which is hard to see from Theorem 2.

As a consequence of Theorem 1, we show that the convergence of the subdivision scheme with a nonnegative mask depends only on the location of its positive coefficients.

Corollary 1. *Let a and c be two nonnegative masks satisfying the sum rule of order one:*

$$\sum_{\alpha\in\mathbb{Z}^s} a(M\alpha+\beta) = \sum_{\alpha\in\mathbb{Z}^s} c(M\alpha+\beta) = 1 \qquad \forall \beta\in\mathbb{Z}^s.$$

Suppose $c(\alpha) > 0$ whenever $a(\alpha) > 0$. If the subdivision scheme associated with mask a converges, then the subdivision scheme associated with mask c also converges.

Proof. Let K be given for the mask c as in Theorem A. Denote F_a and F_c as the transfer matrices defined by (2.2) associated with the masks a and c, respectively. Then for $\alpha, \beta \in K$, $(F_a)_{\alpha,\beta} > 0$ implies that $(F_c)_{\alpha,\beta} > 0$.

Since the subdivision scheme associated with mask a converges, by Theorem 1, for some $1 \le n \le (3^{\#K} - 2^{\#K+1} + 1)/2$, the matrix F_a^n is scrambling. Then the matrix F_c^n is also scrambling. By using Theorem 1 again, we conclude that the subdivision scheme associated with mask c also converges. □

4. Primitive matrices and condition E. In this section we present another approach to the L_2-convergence of subdivision schemes associated with nonnegative masks. This approach is based on primitive matrices and the power involved is at most $(N-1)^2 + 1$, much less than $(3^N - 2^{N+1} + 1)/2$ when N is large.

To state the main result here, we need the *Frobenius normal form* of the column-stochastic matrix F:

$$F = \begin{bmatrix} F_1 & 0 & \cdots & 0 & F_{1,k+1} & \cdots & F_{1,d} \\ 0 & F_2 & 0\cdots & 0 & F_{2,k+1} & \cdots & F_{2,d} \\ \vdots & 0 & \ddots & 0 & \vdots & & \vdots \\ 0 & \cdots & 0 & F_k & F_{k,k+1} & & F_{k,d} \\ 0 & \cdots & 0 & 0 & F_{k+1} & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \ddots & \\ 0 & \cdots & 0 & 0 & 0 & 0 & F_d \end{bmatrix}. \tag{4.1}$$

Here $k \geq 1$ and each F_j is either a 1×1 zero matrix or is irreducible. The blocks $F_1, \ldots, F_k$ are called *isolated*. For each l with $k+1 \leq l \leq d$ (nonisolated block), there exists some i such that $F_{i,l} \neq 0$. The Frobenius normal form for the matrix F can be realized by choosing a suitable order for the set K.

Condition (b) of Theorem A for F is equivalent to condition E. A square matrix A is said to satisfy *condition E* if 1 is a simple eigenvalue of A and all the other eigenvalues are less than 1 in modulus. An irreducible nonnegative matrix A is called *primitive* if A satisfies condition E. Recall that for an irreducible column-stochastic matrix, the primitivity is equivalent to that for some $n \in \mathbb{N}$, A^n is positive (all the entries are positive). The smallest n with this property is called the index of A and can be bounded sharply by $(N-1)^2 + 1$. For these facts on nonnegative matrices, see, e.g., [17].

Using the Frobenius normal form (4.1) we can characterize the L_2-convergence of subdivision schemes with nonnegative masks in terms of primitive matrices.

THEOREM 2. *Let M be a dilation matrix and let $a := \{a(\alpha)\}_{\alpha \in \mathbb{Z}^s}$ be a finitely supported* nonnegative *sequence with $\sum a(\alpha) = m$. Set K and F as in Theorem A. Let $N := \#K$. Assume that the matrix F takes the form* (4.1). *Then the subdivision scheme associated with* (2.1) *converges in L_2 if and only if*

(i) $\sum_{\alpha \in \mathbb{Z}^s} a(M\alpha + \beta) = 1 \; \forall \beta \in \mathbb{Z}^s$;

(ii) *in the form* (4.1), $k = 1$, *and F_1 is primitive.*

Proof. Let us prove the equivalence between condition (b) of Theorem A and condition (ii) of Theorem 2.

Suppose $\rho(F|_V) < 1$. Since F is column-stochastic, it has an eigenvalue 1 with a left eigenvector $[1, \ldots, 1]$. Then F satisfies condition E. Since F is column-stochastic, each of the isolated blocks $F_1, \ldots, F_k$ is also column-stochastic and provides one multiplicity of the eigenvalue 1. Hence, $k = 1$. All the eigenvalues of F_1, including 1, are eigenvalues of F, but F satisfies condition E. Therefore, F_1 satisfies condition E. This in connection with the irreducibility of F_1 implies that F_1 is primitive. Thus condition (ii) of Theorem 2 holds.

Conversely, assume that $k = 1$ and F_1 is primitive. We need to show that $\rho(F_j) < 1$ for $2 \leq j \leq d$.

Suppose to the contrary that for some j with $2 \leq j \leq d$, $\rho(F_j) \geq 1$. Then F_j is not the zero matrix and is irreducible. Since all the eigenvalues of F_j are eigenvalues of F, $\rho(F_j) = 1$. However, F_j is nonnegative. The Perron–Frobenius theorem tells us that F_j has a left positive eigenvector v with eigenvalue 1: $vF_j = v$ and $v = [v_1, \ldots, v_l]$ with $v_1, \ldots, v_l > 0$.

Let $v_{\max} := \max\{v_i : 1 \leq i \leq l\} > 0$ and let $I := \{i : v_i = v_{\max}\} \neq \emptyset$. Then for $i \in I$,

$$v_i = v_{\max} = \sum_{p=1}^{l} v_p (F_j)_{p,i} = v_{\max} \sum_{p \in I} (F_j)_{p,i} + \sum_{p \notin I} v_p (F_j)_{p,i} \leq v_{\max}.$$

Since the equality holds, we know that for $i \in I, p \notin I, (F_j)_{p,i} = 0$. However, F_j is irreducible. Therefore, $I^c = \emptyset$, i.e., $v_1 = v_2 = \cdots = v_l = v_{\max}$. It follows that F_j is column-stochastic. Hence, $F_{i,j} = 0$ for $i \neq j$ and F_j is an isolated block, which is a contradiction.

Thus, we have shown that $\rho(F_j) < 1$ for $2 \leq j \leq d$. But F_1 satisfies condition E. Hence F satisfies condition E, i.e., $\rho(F|_V) < 1$. The proof of Theorem 2 is complete. □

The proof of Theorem 2 yields the following result on general column-stochastic matrices, which is of independent interest.

THEOREM 3. *Let F be column-stochastic and take the Frobenius normal form* (4.1). *Then the following statements are equivalent:*

(i) *F satisfies condition E;*

(ii) *$k = 1$ and F_1 is primitive;*

(iii) *F^n is scrambling for some $n \in \mathbb{N}$;*

(iv) $\rho(F|_V) < 1$.

To apply Theorems 2 and 3, we have to check whether F_1 is primitive. Since F_1 is irreducible, this is equivalent to the fact that F_1^n is scrambling for some $n \in \mathbb{N}$. From the sharp bound $(N-1)^2+1$ for the index of primitive matrices, we know that n can be bounded by $(N-1)^2+1$. The following example shows that, in general, n should be at least $[(N-1)^2/2]+1$.

Example 1. Let A be the following $N \times N$ irreducible column-stochastic matrix:

$$A = \begin{bmatrix} 0 & \cdots & 0 & 1/2 \\ & & & 1/2 \\ & & & 0 \\ & I_{N-1} & & \vdots \\ & & & 0 \end{bmatrix}.$$

Then by computation we see that $A^{[(N-1)^2/2]}$ is not scrambling, while $A^{[(N-1)^2/2]+1}$ is scrambling.

5. Characterizations for self-similar lattice tilings.

In this section we characterize self-similar lattice tilings in terms of finite powers of finite matrices and digit sets.

Observe that the characteristic function $\chi_{T(M,\mathcal{D})}$ of the self-similar tiling (1.1) satisfies the refinement equation

$$\phi(x) = \sum_{\alpha \in \mathcal{D}} \phi(Mx - \alpha) \tag{5.1}$$

with the mask being a $0-1$ sequence (hence, nonnegative). The normalized solution is $\phi = \frac{1}{\text{meas}(T(M,\mathcal{D}))}\chi_{T(M,\mathcal{D})}$. The mask satisfies condition (i) of Theorem 1. Recall from [6] that the symbol of this mask is a conjugate quadrature filter (CQF). Therefore, a well-known fact on CQFs tells us that the integer translates of ϕ are orthonormal (i.e., $T(M,\mathcal{D})$ has measure one) if and only if the subdivision scheme associated with (5.1) converges in L_2. Recall from [7] that for any finitely supported sequence b on $\mathbb{Z}^s$ and any finite set $H \subset \mathbb{Z}^s$, there exists a finite set $K \subset \mathbb{Z}^s$ such that $H \subset K$ and $\ell(K)$ is invariant under $(b(M\alpha - \beta))$. Also, $\ell(K)$ is invariant under $(b(M\alpha - \beta))$ if and only if $M^{-1}(\text{supp} b + K) \cap \mathbb{Z}^s \subset K$. Then we have the following characterization for self-similar lattice tilings.

THEOREM 4. *Let M be a dilation matrix and let $\mathcal{D}$ be a complete set of representatives of distinct cosets of $\mathbb{Z}^s/M\mathbb{Z}^s$. Let $a := \{a(\alpha)\}_{\alpha\in\mathbb{Z}^s}$ be supported on $\mathcal{D}$ and $a(\alpha) = 1$ for $\alpha \in \mathcal{D}$. Let b be the sequence given by*

$$b(\alpha) = \frac{1}{m}\sum_{\beta\in\mathbb{Z}^s} a(\beta)a(\alpha+\beta), \qquad \alpha \in \mathbb{Z}^s.$$

Choose a finite set K such that K contains $\Omega := \mathcal{D}-\mathcal{D}$ and $\ell(K)$ is invariant under the bi-infinite matrix $(b(M\alpha-\beta))$. Let

$$F := \big(b(M\alpha-\beta)\big)_{\alpha,\beta\in K}.$$

Set $N := \#K$. Then the following statements are equivalent:

(1) *$T(M,\mathcal{D})$ is a lattice tiling;*

(2) *there exists some integer n with $1 \le n \le (3^N - 2^{N+1} + 1)/2$ such that the matrix F^n is scrambling;*

(3) *there is no nonempty set $\Lambda \subset K \setminus \Omega$ such that $M^{-1}(\Omega+\Lambda)\cap\mathbb{Z}^s \subset \Lambda$;*

(4) *there is no nonempty finite set $\Lambda \subset \mathbb{Z}^s \setminus \Omega$ such that $M^{-1}(\Omega+\Lambda)\cap\mathbb{Z}^s \subset \Lambda$.*

Proof. The equivalence between statements (1) and (2) follows from Theorem 1.

To see the equivalence between statements (1) and (3), we observe that

$$b(0) = 1.$$

Since F is column-stochastic, it follows that $F_{0,\beta} = 0$ for $\beta \in K\setminus\{0\}$. Then $F_1 = [1]$ is an isolated block in the Frobenius normal form (4.1). By Theorem 2, $T(M,\mathcal{D})$ is a lattice tiling if and only if $[1]$ is the only isolated block in (4.1); that is, there is no nonempty set $\Lambda \subset K\setminus\{0\}$ such that $\ell(\Lambda)$ is invariant under F:

$$F_{\alpha,\beta} > 0, \quad \beta\in\Lambda \qquad \Longrightarrow \qquad \alpha\in\Lambda.$$

Notice that $F_{0,\beta} = b(-\beta) > 0$ for any $\beta\in\Omega$. Therefore, the above statement is equivalent to the fact that there is no nonempty set $\Lambda \subset K\setminus\Omega$ such that

$$F_{\alpha,\beta} = b(M\alpha-\beta) > 0, \quad \beta\in\Lambda \qquad \Longrightarrow \qquad \alpha\in\Lambda,$$

i.e.,

$$M\alpha-\beta\in\Omega, \quad \beta\in\Lambda \qquad \Longrightarrow \qquad \alpha\in\Lambda.$$

This tells us the equivalence between statements (1) and (3). By choosing K to be large enough ($\Lambda \subset K$), we see easily the equivalence between statements (3) and (4). □

Theorem 4 yields a new proof for the first part of the following nice characterization of self-similar tilings. (The second part is new.)

Theorem 5. *Let M be a dilation matrix and let $\mathcal{D}$ be a complete set of representatives of distinct cosets of $\mathbb{Z}^s/M\mathbb{Z}^s$. Define $T(M,\mathcal{D})$ by (1.1). Then $T(M,\mathcal{D})$ is a lattice tiling if and only if*

$$\cup_{n=1}^{\infty}\Big(\sum_{j=0}^{n-1} M^j(\mathcal{D}-\mathcal{D})\Big) = \mathbb{Z}^s. \tag{5.2}$$

Moreover, choose K as in Theorem 4 and let $N = \#K$; then $T(M,\mathcal{D})$ is a lattice tiling if and only if $K \subset \sum_{j=0}^{n-1} M^j(\mathcal{D}-\mathcal{D})$ for some n with $1\le n\le N-m+1$ if and only if $K \subset \sum_{j=0}^{n-1} M^j(\mathcal{D}-\mathcal{D})$ for some $n\in\mathbb{N}$.

Proof. By a translation, we may assume that $0\in\mathcal{D}$. Let a and b be as in Theorem 4. Then the support of b is exactly $\Omega := \mathcal{D}-\mathcal{D}$. Let us prove the first equivalence.

Necessity. Suppose that $T(M,\mathcal{D})$ is a lattice tiling. If (5.2) does not hold, we choose a finite set K such that K contains Ω, $K\cap\big[\mathbb{Z}^s\setminus\cup_{n=1}^{\infty}(\sum_{j=0}^{n-1}M^j\Omega)\big]\ne\emptyset$, and $\ell(K)$ is invariant under $(b(M\alpha-\gamma))$.

Set

$$\Lambda := K \cap \left[\mathbb{Z}^s \setminus \cup_{n=1}^{\infty} \left(\sum_{j=0}^{n-1} M^j \Omega \right) \right].$$

Obviously, $\Lambda \subset K \setminus \Omega$. We state that

$$M^{-1}(\Omega + \Lambda) \cap \mathbb{Z}^s \subset \Lambda.$$

Since $\ell(K)$ is invariant under $(b(M\alpha - \gamma))$, we know that

$$M^{-1}(\Omega + \Lambda) \cap \mathbb{Z}^s \subset M^{-1}(\Omega + K) \cap \mathbb{Z}^s \subset K.$$

Let $\alpha = M^{-1}(\omega + \lambda) \in M^{-1}(\Omega + \Lambda) \cap \mathbb{Z}^s$ with $\omega \in \Omega$ and $\lambda \in \Lambda$. Then $\alpha \in \cup_{j=0}^{n-1} M^j \Omega$ would imply $\lambda = -\omega + M\alpha \in \cup_{j=0}^{n} M^j \Omega$, a contradiction. Therefore, $\alpha \notin \cup_{j=0}^{n-1} M^j \Omega$ for any $n \in \mathbb{N}$. Hence, $\big(M^{-1}(\Omega + \Lambda) \cap \mathbb{Z}^s\big) \cap \big[\cup_{n=1}^{\infty} (\sum_{j=0}^{n-1} M^j \Omega)\big] = \emptyset$. Thus, $M^{-1}(\Omega + \Lambda) \cap \mathbb{Z}^s \subset \Lambda$, and our statement has been proved. This contradicts the condition (4) of Theorem 4. Therefore, (5.2) must be true.

Sufficiency. Suppose (5.2) holds. Choose K as in Theorem 4. Then there exists some $n \in \mathbb{N}$ such that

$$K \subset \sum_{j=0}^{n-1} M^j \Omega. \tag{5.3}$$

Let

$$F := \big(b(M\alpha - \gamma)\big)_{\alpha, \gamma \in K}.$$

We state that

$$\big(F^n\big)_{0,\alpha} > 0 \qquad \forall \alpha \in K.$$

To see this, let $\alpha \in K$. Then (5.3) tells us that there exist $\beta_0, \beta_1, \ldots, \beta_{n-1} \in \Omega$ such that

$$\alpha = \sum_{j=0}^{n-1} M^j \beta_j = \beta_0 + M\beta_1 + \cdots + M^{n-1}\beta_{n-1}.$$

Define

$$\alpha_j := \sum_{i=n-j}^{n-1} M^{i+j-n} \beta_i = \beta_{n-j} + M\beta_{n-j+1} + \cdots + M^{j-1}\beta_{n-1}, \qquad 1 \le j \le n-1.$$

Then

$$M\alpha_{n-1} - \alpha = -\beta_0 \in \Omega \qquad \text{and} \qquad \alpha_1 = \beta_{n-1} \in \Omega.$$

Moreover, for $2 \le j \le n-1$,

$$M\alpha_{j-1} - \alpha_j = \sum_{i=n-j+1}^{n-1} M^{i+j-n} \beta_i - \sum_{i=n-j}^{n-1} M^{i+j-n} \beta_i = -\beta_{n-j} \in \Omega.$$

It follows that $F_{\alpha_{j-1},\alpha_j} = b(M\alpha_{j-1} - \alpha_j) > 0$. Hence,

$$\big(F^n\big)_{0,\alpha} \geq F_{0,\alpha_1}\big(\Pi_{j=2}^{n-1} F_{\alpha_{j-1},\alpha_j}\big) F_{\alpha_{n-1},\alpha} > 0.$$

Thus our statement holds. Hence, F^n is scrambling. By Theorem 4, $T(M, \mathcal{D})$ is a lattice tiling. This proves the sufficiency.

Using the above proof, we can see that $T(M, \mathcal{D})$ is a lattice tiling if $K \subset \sum_{j=0}^{n-1} M^j\Omega$ for some $n \in \mathbb{N}$.

To see the necessity of the second equivalence, choose $1 \leq n \leq 1 + \#(K \setminus \Omega)$ such that

$$K \setminus \sum_{j=0}^{n-1} M^j\Omega = K \setminus \sum_{j=0}^{n} M^j\Omega.$$

Set $\Lambda := K \setminus \sum_{j=0}^{n-1} M^j\Omega \subset K \setminus \Omega \subset K$. Then

$$M^{-1}(\Omega + \Lambda) \cap \mathbb{Z}^s \subset M^{-1}(\Omega + K) \cap \mathbb{Z}^s \subset K.$$

If $\alpha = M^{-1}(\omega + \lambda) \in M^{-1}(\Omega + \Lambda) \cap \mathbb{Z}^s$ with $\omega \in \Omega$ and $\lambda \in \Lambda$, then $\alpha \in \sum_{j=0}^{n-1} M^j\Omega$ would imply $\lambda = -\omega + M\alpha \in \cup_{j=0}^{n} M^j\Omega$, a contradiction. Therefore,

$$\big(M^{-1}(\Omega + \Lambda) \cap \mathbb{Z}^s\big) \cap \left[\sum_{j=0}^{n-1} M^j\Omega\right] = \emptyset.$$

Hence, $M^{-1}(\Omega + \Lambda) \cap \mathbb{Z}^s \subset \Lambda$. Since $T(M, \mathcal{D})$ is a lattice tiling, by Theorem 4, $\Lambda = \emptyset$, i.e., $K \subset \sum_{j=0}^{n-1} M^j\Omega$. Since $\#(K \setminus \Omega) \leq N - m$, the necessity of the second statement is proved. □

As a corollary we obtain another proof of the necessary condition due to Lagarias and Wang [10].

Corollary 2. *Let M be a dilation matrix and let $\mathcal{D}$ be a complete set of representatives of distinct cosets of $\mathbb{Z}^s/M\mathbb{Z}^s$. Define $T(M, \mathcal{D})$ by (1.1). If $T(M, \mathcal{D})$ is a lattice tiling, then the smallest M-invariant sublattice $\mathbb{Z}[M, \mathcal{D}]$ of $\mathbb{Z}^s$ containing the difference set $\mathcal{D} - \mathcal{D}$ is $\mathbb{Z}^s$.*

Let us give an example to show the difference between the necessary condition in [10] and the necessary and sufficient conditions in Theorem 5.

Example 2. Let M and $\mathcal{D}$ be given as

$$M := \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}, \qquad \mathcal{D} := \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 3 \\ 1 \end{bmatrix} \right\}.$$

Then $\mathbb{Z}[M, \mathcal{D}] = \mathbb{Z}^s$ by [10]. However,

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \notin \cup_{n=1}^{\infty} \left(\sum_{j=0}^{n-1} M^j\Omega \right).$$

Hence, $T(M, \mathcal{D})$ is not a lattice tiling.

Proof. Notice that

$$\Omega = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \begin{bmatrix} -3 \\ 1 \end{bmatrix}, \begin{bmatrix} -3 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \begin{bmatrix} -3 \\ -1 \end{bmatrix}, \begin{bmatrix} 3 \\ -1 \end{bmatrix} \right\}.$$

Then the second component of each vector in $M^j\Omega$ is in the set $\{0, 2^j, -2^j\}$.

Suppose that for some $n \in \mathbb{N}$ and $\alpha_0, \ldots, \alpha_{n-1} \in \Omega$,

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \sum_{j=0}^{n-1} M^j \alpha_j.$$

Comparing the second components, we see that the second components of $\alpha_0, \ldots, \alpha_{n-1}$ are all zero. But the only vectors with this property in $M^j\Omega$ are in $2^j\Omega_0$, where

$$\Omega_0 = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 0 \end{bmatrix}, \begin{bmatrix} -3 \\ 0 \end{bmatrix} \right\}.$$

Therefore, the first component of $\sum_{j=0}^{n-1} M^j \alpha_j$ lies in $3\mathbb{Z}$, which is a contradiction. Thus, the condition of Theorem 5 does not hold, and $T(M, \mathcal{D})$ is not a lattice tiling. □

The condition $K \subset \sum_{j=0}^{n-1} M^j(\mathcal{D} - \mathcal{D})$ can be reduced into the *knapsack problem*, which can be solved by a polynomial-time algorithm for the fixed n (see, e.g., [14]). The necessity condition $\mathbb{Z}[M, \mathcal{D}] = \mathbb{Z}^s$ can be easily checked by the solvability of linear diophantine equations.

6. Examples. In this section we provide some examples to show the applicability of the characterizations stated in Theorems 4 and 5.

Example 3. Let M be the dilation matrix

$$M = \begin{bmatrix} a & b \\ 0 & c \end{bmatrix}$$

with $a > 1, c > 1$. Choose $\{k_0 = 0, k_1, \ldots, k_{a-1}\} \subset \mathbb{Z}$ and $\{p_0 = 0, p_1, \ldots, p_{c-1}\} \subset \mathbb{Z}$ such that $k_j \equiv j(\mathrm{mod} a)$ and $p_i \equiv i(\mathrm{mod} c)$. Let $l_0, l_1, \ldots, l_{c-1}$ be arbitrary integers and set

$$\mathcal{D} = \left\{ \begin{bmatrix} l_i + k_j \\ p_i \end{bmatrix} : \quad i = 0, 1, \ldots, c-1; \, j = 0, 1, \ldots, a-1 \right\}.$$

Then $\mathcal{D}$ is a complete set of representatives of distinct cosets of $\mathbb{Z}^2/M\mathbb{Z}^2$. The self-similar tiling $T(M, \mathcal{D})$ is a lattice tiling if and only if $\gcd(k_1, \ldots, k_{a-1}) = 1$ and $\gcd(p_1, \ldots, p_{c-1}) = 1$.

Proof. Let $\varepsilon_1 = (l_{i_1} + k_{j_1}, p_{i_1})^T, \varepsilon_2 = (l_{i_2} + k_{j_2}, p_{i_2})^T \in \mathcal{D}$. Suppose $\varepsilon_1 - \varepsilon_2 \in M\mathbb{Z}^2$. Then for some $\beta = (\beta_1, \beta_2)^T \in \mathbb{Z}^2$,

$$\varepsilon_1 - \varepsilon_2 = \begin{bmatrix} l_{i_1} + k_{j_1} - l_{i_2} - k_{j_2} \\ p_{i_1} - p_{i_2} \end{bmatrix} = M\beta = \begin{bmatrix} a\beta_1 + b\beta_2 \\ c\beta_2 \end{bmatrix}.$$

Both i_1 and i_2 are in $\{0, 1, \ldots, c-1\}$ and $p_i \equiv i(\mathrm{mod} c)$. Hence, $\beta_2 = 0$ and $i_1 = i_2$. It follows that $k_{j_1} - k_{j_2} = a\beta_1 \in a\mathbb{Z}$. But $k_j \equiv j(\mathrm{mod} a)$. We know that $j_1 = j_2$, i.e., $\varepsilon_1 = \varepsilon_2$. This proves that $\mathcal{D}$ is a complete set of representatives of distinct cosets of $\mathbb{Z}^2/M\mathbb{Z}^2$.

To see the second statement, first prove the sufficiency. Suppose $\gcd(k_1, \ldots, k_{a-1}) = 1$ and $\gcd(p_1, \ldots, p_{c-1}) = 1$. We show that $\mathbb{Z}^2 \subset \cup_{n=1}^{\infty} \sum_{j=0}^{n-1} M^j(\mathcal{D} - \mathcal{D})$. Then by Theorem 5, $T(M, \mathcal{D})$ is a lattice tiling.

Let $\alpha = (\alpha_1, \alpha_2)^T \in \mathbb{Z}^2$. Theorem 5 in connection with the characterization of univariate self-similar lattice tilings in [5, 21] (see also the introduction) tells us that there are $n \in \mathbb{N}$ and $i_0, i_1, \ldots, i_{n-1}, \tilde{i}_0, \tilde{i}_1, \ldots, \tilde{i}_{n-1} \in \{0, 1, \ldots, c-1\}$ such that

$$\alpha_2 = \sum_{j=0}^{n-1} c^j (p_{i_j} - p_{\tilde{i}_j}).$$

Then in the same way, there are $N \in \mathbb{N}$ and $\{n_j, \tilde{n}_j\}_{j=0}^{N-1} \subset \{0, 1, \ldots, a-1\}$ such that

$$\sum_{j=0}^{N-1} a^j (k_{n_j} - k_{\tilde{n}_j}) = \alpha_1 - \sum_{j=0}^{n-1} a^j (l_{i_j} - l_{\tilde{i}_j}) - \sum_{j=0}^{n-1} b(p_{i_j} - p_{\tilde{i}_j}) \sum_{q=0}^{j-1} a^q c^{j-1-q}.$$

We may assume that $N \geq n$, since otherwise we need only to choose $n_j = \tilde{n}_j$ for $j = N, \ldots, n-1$.

Observe that

$$M^j = \begin{bmatrix} a^j & b\sum_{q=0}^{j-1} a^q c^{j-1-q} \\ 0 & c^j \end{bmatrix}.$$

Take $i_j = \tilde{i}_j$ for $j = n, \ldots, N-1$. Then

$$\sum_{j=0}^{N-1} M^j \left(\begin{bmatrix} l_{i_j} + k_{n_j} \\ p_{i_j} \end{bmatrix} - \begin{bmatrix} l_{\tilde{i}_j} + k_{\tilde{n}_j} \\ p_{\tilde{i}_j} \end{bmatrix} \right)$$
$$= \begin{bmatrix} \sum_{j=0}^{N-1} a^j (k_{n_j} - k_{\tilde{n}_j}) + \sum_{j=0}^{n-1} a^j (l_{i_j} - l_{\tilde{i}_j}) + \sum_{j=0}^{n-1} b(p_{i_j} - p_{\tilde{i}_j}) \sum_{q=0}^{j-1} a^q c^{j-1-q} \\ \sum_{j=0}^{n-1} c^j (p_{i_j} - p_{\tilde{i}_j}) \end{bmatrix}.$$

The first component of this vector is exactly α_1, while the second is α_2. Therefore, $\alpha \in \sum_{j=0}^{N-1} M^j(\mathcal{D} - \mathcal{D})$. Hence, $\mathbb{Z}^2 \subset \cup_{n=1}^{\infty} \sum_{j=0}^{n-1} M^j(\mathcal{D} - \mathcal{D})$.

To see the necessity, suppose $T(M, \mathcal{D})$ is a lattice tiling. Then by Theorem 5, $\mathbb{Z}^2 \subset \cup_{n=1}^{\infty} \sum_{j=0}^{n-1} M^j(\mathcal{D} - \mathcal{D})$.

Let $N \in \mathbb{N}$ and $\{i_j, \tilde{i}_j\}_{j=0}^{N-1} \subset \{0, 1, \ldots, c-1\}, \{n_j, \tilde{n}_j\}_{j=0}^{N-1} \subset \{0, 1, \ldots, a-1\}$. Then

$$\begin{aligned} &\sum_{j=0}^{N-1} M^j \left(\begin{bmatrix} l_{i_j} + k_{n_j} \\ p_{i_j} \end{bmatrix} - \begin{bmatrix} l_{\tilde{i}_j} + k_{\tilde{n}_j} \\ p_{\tilde{i}_j} \end{bmatrix} \right) \\ &= \begin{bmatrix} \sum_{j=0}^{N-1} a^j (l_{i_j} - l_{\tilde{i}_j} + k_{n_j} - k_{\tilde{n}_j}) + \sum_{j=0}^{N-1} b(p_{i_j} - p_{\tilde{i}_j}) \sum_{q=0}^{j-1} a^q c^{j-1-q} \\ \sum_{j=0}^{n-1} c^j (p_{i_j} - p_{\tilde{i}_j}) \end{bmatrix}. \end{aligned} \tag{6.1}$$

The second component of (6.1) is divisible by $\gcd(p_1, \ldots, p_{c-1})$. Therefore, there must hold $\gcd(p_1, \ldots, p_{c-1}) = 1$.

We also know that $(1, 0)^T \in \sum_{j=0}^{N-1} M^j(\mathcal{D} - \mathcal{D})$ for some N. Suppose the vector (6.1) equals $(1, 0)^T$. Then

$$\sum_{j=0}^{N-1} c^j (p_{i_j} - p_{\tilde{i}_j}) = 0.$$

Hence, $p_{i_0} - p_{\tilde{i}_0} \in c\mathbb{Z}$. But $p_i \equiv i (\mathrm{mod} c)$. Therefore, $i_0 = \tilde{i}_0$. In the same way, $i_1 = \tilde{i}_1, \ldots, i_{N-1} = \tilde{i}_{N-1}$.

Consider the first component of (6.1). It states that

$$1=\sum_{j=0}^{N-1} a^j (k_{n_j}-k_{\tilde{n}_j}).$$

This is divisible by $\gcd(k_1,\dots,k_{a-1})$. Hence, $\gcd(k_1,\dots,k_{a-1})=1$. □

If we take $a=c=2, b=0, p_1=1, k_1=3$, and $l_0=l_1=0$, we obtain Example 2. The case $p_i=i, k_j=j$ was proved in [5].

Let us turn to the necessary condition $\mathbb{Z}[M,\mathcal{D}]=\mathbb{Z}^s$ in our special example. Assume $l_0, = l_1=\cdots=l_{c-1}=0$.

Example 4. Let $M,\mathcal{D}$ as in Example 3 and $l_0=l_1=\cdots=l_{c-1}=0$. Then $\mathbb{Z}[M,\mathcal{D}]=\mathbb{Z}^2$ if and only if $\gcd(p_1,\dots,p_{c-1})=1$ and $\gcd(b,k_1,\dots,k_{a-1})=1$.

Proof. Since $0\in\mathcal{D}, \mathbb{Z}[M,\mathcal{D}]=\mathbb{Z}[\mathcal{D},M\mathcal{D}]$ as shown in [10]. Let Δ be a $2\times(ac)$ matrix whose columns are vectors in $\mathcal{D}$. Then the solvability of linear diophantine equations tells us that $\mathbb{Z}[\mathcal{D},M\mathcal{D}]=\mathbb{Z}^2$ if and only if the greatest common divisor of 2×2 minors of the matrix $[\Delta, M\Delta]$ is one. Note that the columns of $M\Delta$ have the form

$$M\begin{bmatrix} k_j \\ p_i \end{bmatrix} = \begin{bmatrix} ak_j+bp_i \\ cp_i \end{bmatrix}.$$

Each entry of the first row of the matrix $[\Delta, M\Delta]$ is divisible by $\gcd(b,k_1,\dots,k_{a-1}) = 1$, while each entry of the second row is divisible by $\gcd(p_1,\dots,p_{c-1})=1$. The necessity follows easily.

Conversely, if $\gcd(p_1,\dots,p_{c-1})=\gcd(b,k_1,\dots,k_{a-1})=1$, we choose the following 2×2 minors of the matrix $[\Delta, M\Delta]$:

$$\begin{vmatrix} k_j & 0 \\ 0 & p_i \end{vmatrix} = k_j p_i, \qquad \begin{vmatrix} bp_i & 0 \\ cp_i & p_{\tilde{i}} \end{vmatrix} = bp_i p_{\tilde{i}}.$$

The greatest common divisor of these minors is

$$\gcd\bigl(\gcd(k_1,\dots,k_{a-1})\cdot\gcd(p_1,\dots,p_{c-1}), \quad b\cdot[\gcd(p_1,\dots,p_{c-1})]^2\bigr),$$

which equals one. Therefore, the greatest common divisor of 2×2 minors of the matrix $[\Delta, M\Delta]$ is one, and $\mathbb{Z}[M,\mathcal{D}]=\mathbb{Z}^2$. □

Example 5. Let $2\le m\in\mathbb{N}$ and let $M=mI_s$. If $\mathcal{D}$ is a complete set of representatives of distinct cosets of $\mathbb{Z}^s/M\mathbb{Z}^s$ and

$$\Bigl\{0,1,\dots,\Bigl[\frac{m}{2}\Bigr]\Bigr\}^s\subset\mathcal{D},$$

then $T(M,\mathcal{D})$ is a lattice tiling.

Proof. By our assumption,

$$\Bigl\{-\Bigl[\frac{m}{2}\Bigr],\dots,-1,0,1,\dots,\Bigl[\frac{m}{2}\Bigr]\Bigr\}^s\subset\mathcal{D}-\mathcal{D}.$$

Then for $n\in\mathbb{N}$,

$$\Bigl[-\Bigl[\frac{m}{2}\Bigr]\frac{m^n-1}{m-1}, \quad \Bigl[\frac{m}{2}\Bigr]\frac{m^n-1}{m-1}\Bigr]^s\cap\mathbb{Z}^s\subset\sum_{j=0}^{n-1}M^j(\mathcal{D}-\mathcal{D}).$$

Therefore, for sufficiently large n,

$$K \subset \sum_{j=0}^{n-1} M^j(\mathcal{D} - \mathcal{D}).$$

By Theorem 5, $T(M, \mathcal{D})$ is a lattice tiling. $\square$

In particular, for $M = \left[\begin{smallmatrix}3 & 0\\ 0 & 3\end{smallmatrix}\right]$, if $\mathcal{D}$ contains $(0,0)^T, (1,0)^T, (0,1)^T, (1,1)^T$, then $T(M, \mathcal{D})$ is a lattice tiling.

REFERENCES

[1] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary Subdivision*, Mem. Amer. Math. Soc. 93, 1991.

[2] 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.

[3] I. Daubechies and J. C. Lagarias, *Sets of all finite products of which converge*, Linear Algebra Appl., 161 (1992), pp. 227–263.

[4] T. N. T. Goodman, C. A. Micchelli, and J. D. Ward, *Spectral radius formulas for subdivision operators*, in Recent Advances in Wavelet Analysis, L. L. Schumaker and G. Webb, eds., Wavelet Anal. Appl. 3, Academic Press, Boston, 1994, pp. 335–360.

[5] K. Gröchenig and A. Haas, *Self-similar lattice tilings*, J. Fourier Anal. Appl., 1 (1994), pp. 131–170.

[6] K. Gröchenig and W. Madych, *Multiresolution analysis, Haar bases, and self-similar tilings*, IEEE Trans. Inform. Theory, 38 (1992), pp. 556–568.

[7] B. Han and R.-Q. Jia, *Multivariate refinement equations and convergence of subdivision schemes*, SIAM J. Math. Anal., 29 (1998), pp. 1177–1199.

[8] R. Q. Jia, *Subdivision schemes in L_p spaces*, Adv. Comput. Math., 3 (1995), pp. 309–341.

[9] R.-Q. Jia and D.-X. Zhou, *Convergence of subdivision schemes associated with nonnegative masks*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 418–430.

[10] J. C. Lagarias and Y. Wang, *Integral self-affine tiles in $\mathbb{R}^n$.* I. *Standard and nonstandard digit sets*, J. London Math. Soc. (2), 54 (1996), pp. 161–179.

[11] J. C. Lagarias and Y. Wang, *Self-affine tiles in $\mathbb{R}^n$*, Adv. Math., 121 (1996), pp. 21–49.

[12] W. Lawton, *Necessary and sufficient conditions for constructing orthonormal wavelet bases*, J. Math. Phys., 32 (1991), pp. 57–61.

[13] W. Lawton, S. L. Lee, and Z. W. Shen, *Convergence of multidimensional cascade algorithms*, Numer. Math., 78 (1998), pp. 427–438.

[14] H. W. Lenstra, Jr., *Integer programming and cryptography*, Math. Intelligencer, 6 (1984), pp. 14–19.

[15] C. A. Micchelli and H. Prautzsch, *Refinement and subdivision for spaces of integer translates of a compactly supported function*, in Numerical Analysis 1987, D. F. Griffiths and G. A. Watson, eds., Pitman Res. Notes Math. Ser. 170, Longman Sci. Tech., Harlow, 1988, pp. 192–222.

[16] A. Paz, *Definite and quasidefinite sets of stochastic matrices*, Proc. Amer. Math. Soc., 16 (1965), pp. 634–641.

[17] E. Seneta, *Non-Negative Matrices*, Wiley, New York, 1973.

[18] G. Strang, *Eigenvalues of $(\downarrow 2)H$ and convergence of cascade algorithm*, IEEE Trans. Signal Process., 44 (1996), pp. 233–238.

[19] G. Strang and D. X. Zhou, *Inhomogeneous refinement equations*, J. Fourier Anal. Appl., 4 (1998), pp. 733–747.

[20] R. S. Strichartz, *Self-similar measures and their Fourier transforms*, Indiana Univ. Math. J., 39 (1990), pp. 797–817.

[21] D.-X. Zhou, *Stability of refinable functions, multiresolution analysis, and Haar bases*, SIAM J. Math. Anal., 27 (1996), pp. 891–904.

[22] D. X. Zhou, *The p-norm joint spectral radius for even integers*, Methods Appl. Anal., 5 (1998), pp. 39–54.

SIAM J. Math. Anal.
Vol. 33, No. 1, pp. 16–52

PARAMETRICALLY EXCITED HAMILTONIAN PARTIAL DIFFERENTIAL EQUATIONS[*]

E. KIRR[†] AND M. I. WEINSTEIN[‡]

Abstract. Consider a linear autonomous Hamiltonian system with a time-periodic bound state solution. In this paper we study the structural instability of this bound state relative to time almost periodic perturbations which are small, localized, and Hamiltonian. This class of perturbations includes those whose time dependence is periodic but encompasses a large class of those with finite (quasi-periodic) or infinitely many noncommensurate frequencies. Problems of the type considered arise in many areas of applications including ionization physics and the propagation of light in optical fibers in the presence of defects. The mechanism of instability is radiation damping due to resonant coupling of the bound state to the continuum modes by the time-dependent perturbation. This results in a transfer of energy from the discrete modes to the continuum. The rate of decay of solutions is slow and hence the decaying bound states can be viewed as metastable. These results generalize those of A. Soffer and M. I. Weinstein, who treated localized time-periodic perturbations of a particular form. In the present work, new analytical issues need to be addressed in view of (i) the presence of infinitely many frequencies which may resonate with the continuum as well as (ii) the possible accumulation of such resonances in the continuous spectrum. The theory is applied to a general class of Schrödinger operators.

Key words. Hamiltonian partial differential equations, parametric resonance, time-dependent perturbation theory, Fermi golden rule, energy transfer, metastable states

AMS subject classifications. 37L50, 35B34, 35B40, 37K55, 35B35, 35P25, 11K70

PII. S0036141099363456

1. Introduction.

1.1. Overview.

Consider a dynamical system of the form

$$i\partial_t\phi \;=\; H_0\;\phi, \tag{1.1}$$

where H_0 denotes a self-adjoint operator on a Hilbert space $\mathcal{H}$. We further assume that H_0 has only one eigenstate $\psi_0 \in \mathcal{H}$ with corresponding simple eigenvalue λ_0. Thus,

$$b_*(t) \;=\; e^{-i\lambda_0 t}\psi_0 \tag{1.2}$$

is a time-periodic *bound state* solution of the dynamical system (1.1). We next introduce the perturbed dynamical system

$$i\partial_t\phi \;=\; (H_0 \;+\; \varepsilon W(t))\,\phi. \tag{1.3}$$

In this paper we prove that if the perturbation, $\varepsilon W(t)$, is small, "generic," and almost periodic in time,[1] then solutions of the perturbed dynamical system (1.3) tend to zero

[*] Received by the editors October 26, 1999; accepted for publication (in revised form) October 26, 2000; published electronically April 12, 2001. This research was supported in part by National Science Foundation grant DMS-950097.

http://www.siam.org/journals/sima/33-1/36345.html

[†] Department of Mathematics, University of Michigan, Ann Arbor, MI and Department of Applied Mathematics, University Babeş-Bolyai, Cluj, Romania (ekirr@math.lsa.umich.edu). Part of this work was done while this author participated in the Bell Labs/Lucent Student Intern Program.

[‡] Mathematical Sciences Research, Bell Laboratories - Lucent Technologies, Murray Hill, NJ 07974 (miw@research.bell-labs.com).

[1] See the appendix in section 9 as well as [2, 9] for definitions and results on almost periodic functions.

as $t \to \pm\infty$. It follows that the state, $b_*(t)$, does not continue or deform to a time-periodic or even time-almost-periodic state. Thus, $b_*(t)$ is structurally unstable with respect to this class of perturbations. Our methods yield a detailed description of the transient (t large but finite) and long time ($t \to \pm\infty$) behavior solutions to the initial value problem. Theorems 2.1–2.3 contain precise statements of our main results. The following picture emerges concerning time evolution (1.3) for initial data given by the bound state, ψ_0, of the unperturbed problem. Let

$$P(t) = |(\ \psi_0, \phi(t)\ )|^2 \tag{1.4}$$

be the modulus square of the projection of the solution at time t onto the state ψ_0.[2] Then,

(i) $P(t)\ \sim\ 1 - C_W\ |t|^2$ for $|t|$ small,[3]

(ii) $P(t)\ \sim\ \exp(-2\varepsilon^2\Gamma t)$ for $t \le \mathcal{O}((\varepsilon^2\Gamma)^{-1}),\ \Gamma\ = \mathcal{O}(W^2)$, and

(iii) $P(t)\ \sim\ \langle t\rangle^{-\alpha}$ for $|t| >> (\varepsilon^2\Gamma)^{-1}$ for some $\alpha > 0$.

The time $\tau\ =\ (\varepsilon^2\Gamma)^{-1}$ is called the *lifetime* of the state $b_*(t)$, which can be thought of as being *metastable* due to its slow decay. The mechanism for large time decay is resonant coupling of the bound state with continuous spectrum due to the time-dependent perturbation. Our analysis makes explicit the slow transfer of energy from the discrete to continuum modes and the accompanying radiation of energy out of any compact set.

Phenomena of the type considered here are of importance in many areas of theoretical physics and applications. Examples include ionization physics [3, 4, 10] and the propagation of light in optical fibers in the presence of defects [13]; see the discussion below.

The results of this article generalize those of Soffer and Weinstein [22], where the case

$$W(t)\ =\ \cos(\mu t)\ \beta,\quad \beta = \beta^* \tag{1.5}$$

was considered. The method used is a time-dependent/dynamical systems approach introduced in [21], [23] in a perturbation theory of operators with embedded eigenvalues in their continuous spectrum. These ideas were also used in a study of resonant radiation damping of nonlinear systems [24], as well as in a class of parametric resonance problems [22]; see also [12]. New analytical questions must be addressed in view of (i) the presence of infinitely many frequencies which may resonate with the continuum as well as (ii) the possible accumulation of such resonances in the continuous spectrum. This leads to a careful use of almost periodic properties of the perturbation (Theorems 2.1 and 2.2) and hypothesis **(H6)** (Theorem 2.3), which is easily seen to hold when the perturbation, $W(t)$, consists of a sum over a finite number of frequencies, μ_j.

A special case for which the hypotheses of our theorems are verified is the case of the Schrödinger operator $H_0\ =\ -\Delta\ +\ V(x)$. Here, $V(x)$ is a real-valued function of $x \in \mathbb{R}^3$ which decays sufficiently rapidly as $|x| \to \infty$. In this setting Soffer and Weinstein [22] studied in detail the structural instability of $b_*(t)$ by considering the perturbed dynamical system (1.3) with $W(x,t)\ =\ \beta(x)\ \cos(\mu t)$. Here, we consider

[2] (f, g) denotes the inner product of f and g. If ψ_0 is normalized, then $P(t)$ has the quantum mechanical interpretation of the probability that the system at time t is in the state ψ_0.

[3] We do not discuss the short time behavior in this article; see [12]. This small time behavior is related to the “watched pot” effect in quantum measurement theory [15].

a class of perturbations of the form $W(x,t) \;=\; \sum_j \; \beta_j(x) \;\; \cos \mu_j t$, where the sum may be finite or infinite and where the frequencies μ_j need not be commensurate, e.g., $W(x,t) \;=\; \beta_1(x) \;\; \cos t \;+\; \beta_2(x) \;\; \cos \sqrt{2}t$, where $\beta_i(x),\;\; i = 1,2$, is rapidly decaying as $x \to \infty$.

In addition to the problem of ionization by general time-varying fields, we mention other motivations for considering the class of time-dependent perturbations sketched above and defined in detail in section 2.

(a) An area of application to which our analysis applies is the propagation of light through an optical fiber [13]. In the regime where backscattering can be neglected, the propagation of waves down the length of the fiber is governed by a Schrödinger equation:

$$i\partial_z \phi \;=\; (\; -\Delta_\perp \;+\; V(x_\perp))\,\phi \;+\; W(x_\perp, z)\phi. \tag{1.6}$$

Here, ϕ denotes the slowly varying envelope of the highly oscillatory electric field, a function of z, the direction of propagation along the fiber, and $x_\perp \in \mathbb{R}^2$, the transverse variables. $V(x_\perp)$ denotes an unperturbed index of refraction profile and $W(x_\perp, z)$ the small fluctuations in refractive index along the fiber. These can arise due to defects introduced either accidentally or by design. The models considered allow for distributions of defects which are far more general than periodic. Our analysis addresses the simple situation of energy in a single transverse mode propagating and being radiated away due to coupling by defects to continuum modes. The bound state channel sees an effective damping. In particular the results of this paper have been applied to a study of structural instability of so-called breather modes of planar "soliton wave guides" [12]. The case of multiple transverse modes is of great interest [13]. Here, one has the phenomena of coupling among discrete modes as well as the coupling of discrete to continuum/radiation modes [7]. There is extensive interesting work on this problem in the case where $W(x_\perp, z)$ is a stochastic process in z and radiation is neglected [8].

(b) Nonlinear problems can be viewed as linear time-dependent potential problems where the time-dependent potential is given by the solution. *A priori* one knows little about the time dependence of the solution of a nonlinear problem. Nonlinearity is expected, in general, to excite infinitely many frequencies. Therefore results of a general nature for potentials with very general time dependence are of interest. This point of view is adopted by Sigal [19, 20], who considers the case where the nonlinear term defines a time-periodic perturbation and then proceeds to study the resonance problem via time-independent Floquet analysis applied to the so-called Floquet Hamiltonian. The dilation analytic techniques used were first applied in the context of time-periodic Hamiltonians by Yajima [26, 27, 28]. Floquet-type methods were also used in the time-periodic context by Vainberg [25]. The general class of perturbations we consider are not treatable by Floquet analysis and time-dependent analysis appears necessary.

1.2. Outline of the method. We now give a brief outline of our approach. For simplicity consider the initial value problem

$$i\partial_t \phi(t,x) \;=\; H_0 \;\phi(t,x) \;+\; \varepsilon W(t,x)\;\phi(t,x), \tag{1.7}$$

$$\phi|_{t=0} \;=\; \phi(0), \tag{1.8}$$

where

$$H_0 = -\Delta \; + V(x),\;\; W(t,x) = g(t)\;\beta(x),\;\; g(t) = \sum_j \; g_j \; e^{-i\mu_j t} \tag{1.9}$$

is a real-valued almost periodic function of t, and $\beta(x)$ is a real-valued and rapidly decaying function of x as $|x| \to \infty$. The unperturbed problem ($\varepsilon = 0$) can be trivally written as two decoupled equations governing the bound state amplitude, $a(t)$, and dispersive components, $\phi_d(t)$, of the solution. Specifically, let

$$\phi(t) \ = \ a(t)\ \psi_0(x) \ + \ \phi_d(t,x), \ (\psi_0, \phi_d(t)) \ = \ 0. \tag{1.10}$$

Then,

$$\begin{aligned} i\partial_t a(t) &= \lambda_0 a(t), \\ i\partial_t \ \phi_d(t,x) &= H_0 \ \phi_d(t,x) \end{aligned} \tag{1.11}$$

with initial conditions

$$\begin{aligned} a(0) &= (\psi_0, \phi(0))\,, \\ \phi_d(0) &= \mathbf{P_c}\phi(0), \end{aligned} \tag{1.12}$$

where

$$\mathbf{P_c} f \equiv \ f - (\psi_0, f)\, \psi_0$$

defines the projection onto the continuous spectral part of H_0.

For initial data $a(0) = 1$, $\phi_d(0) = 0$, we have $a(t) = e^{-i\lambda_0 t}$, $\phi_d(t) \equiv 0$, corresponding to the bound state, $b_*(t)$.

We now ask the following:

(a) *Under the small perturbation $\varepsilon W(t,x)$, does the bound state deform or continue to a nearby periodic or even almost periodic solution?*

(b) *How do solutions to the perturbed initial value problem behave as $|t| \to \infty$?*

For small perturbations $\varepsilon W(t,x)$ it is natural to use the decomposition (1.10). Substitution of (1.10) into (1.3) yields a weakly coupled system for $a(t)$ and $\phi_d(t)$. This system is derived and analyzed in detail in sections 4–6.

In order to illustrate the main idea, we introduce a simplified system having the same general character:

$$\begin{aligned} i\partial_t a(t) \ &= \ \lambda_0 \ a(t) \ + \ \varepsilon g(t) \ (\beta\psi_0, \phi_d(t)) \\ i\partial_t \phi_d(t,x) \ &= \ -\Delta\phi_d(t,x) \ + \ \varepsilon a(t) g(t) \beta(x) \psi_0(x). \end{aligned} \tag{1.13}$$

Here, we have replaced H_0 on its continuous spectral part by $-\Delta$.

If $\varepsilon\beta$ is small, then $A(t) \equiv e^{i\lambda_0 t} a(t)$ is slowly varying ($\partial_t A(t) = \mathcal{O}(\varepsilon\beta)$). In particular, we have

$$\begin{aligned} i\partial_t A(t) \ &= \ \varepsilon e^{i\lambda_0 t} \ g(t) \ (\beta\psi_0, \phi_d(t)) \\ i\partial_t \phi_d(t,x) \ &= \ -\Delta\phi_d(t,x) \ + A(t) e^{-i\lambda_0 t} \ \varepsilon g(t) \beta(x) \psi_0(x). \end{aligned} \tag{1.14}$$

Viewing $A(t)$ as nearly constant, we see that the inhomogeneous source term in (1.14) has frequencies $\lambda_0 + \mu_j$; see (1.9). Therefore, if $\lambda_0 + \mu_j > 0$ for some j, then $\lambda_0 + \mu_j$ lies in the continuous spectrum of $-\Delta$ (H_0) and therefore ϕ_d satisfies a resonantly forced wave equation. A careful expansion and analysis to second order in the perturbation $\varepsilon W(t)$ (see the proof of Proposition 4.1) reveals the system for $A(t)$ and $\phi_d(t)$ can be rewritten in the following form, in which the effect of this resonance is made explicit:

$$\partial_t A(t) \ = \ (-\varepsilon^2 \Gamma \ + \ \rho(t))\ A(t) \ + \ E(t; A(t), \phi_d(t)), \tag{1.15}$$

$$i\partial_t \phi_d(t,x) \ = \ H_0 \ \phi_d(t,x) \ + \ \mathbf{P_c} \ F(t,x; A(t), \phi_d(t)). \tag{1.16}$$

The terms $E(t)$ and $F(t,x)$ formally tend to zero if $A(t)$ tends to zero and if the "local energy" of $\phi_d(t)$ tends to zero as $t \to \infty$. The strategy of sections 5 and 6 is to derive coupled estimates for $A(t)$ and a measure of the local energy of ϕ_d from which one can conclude, for $\varepsilon W(t)$ small, that solutions to (1.15)–(1.16) decay in an appropriate sense. The key to the decay of solutions is the constant Γ, given by

$$\Gamma \;\equiv\; \frac{\pi}{4} \sum_{\{j\ :\ \lambda_0+\mu_j>0\}} |g_j|^2\ \left(\mathbf{P_c}\beta\psi_0,\ \delta(H_0-\lambda_0-\mu_j)\mathbf{P_c}\beta\psi_0\right); \tag{1.17}$$

see also hypothesis **(H5)** of section 2. The quantity Γ is a generalization of the well-known Fermi golden rule arising in the theory of radiative transitions in quantum mechanics [3, 4, 10]. For the example at hand, (1.9), the sum in (1.17) is over all j for which $\mu_j+\lambda_0$ is strictly positive, i.e., lies in the continuous spectrum of H_0. Thinking of H_0 as having a spectral decomposition in terms of eigenfunctions and generalized eigenfunctions, let $e(\lambda)$ denote a generalized eigenfunction associated with the energy λ. Then each term in the sum (1.17) is of the form

$$|(\ e(\lambda_0+\mu_j), \beta\psi_0\ )|^2\,. \tag{1.18}$$

Thus, clearly $\Gamma > 0$, generically.

Neglecting for the moment the oscillatory function $\rho(t)$ in (1.15), we see that coupling of the bound state by the time dependent perturbation to the continuum-radiation modes, at the frequencies $\mu_j+\lambda_0>0$, leads to decay of the bound state. The leading order of (1.15)–(1.16) is a normal form in which this internal damping effect is made explicit; energy is transferred from the discrete to the continuous spectral components of the solution while the total energy remains independent of time:

$$\begin{aligned} \|\ \phi(t)\ \|_2^2 \;&=\; |a(t)|^2 \;+\; \|\ \phi_d(t)\ \|_2^2 \\ &=\; |a(0)|^2 \;+\; \|\ \phi_d(0)\ \|_2^2. \end{aligned} \tag{1.19}$$

1.3. Energy flow; contrast with the analysis of [22]. The goal is to show that energy flows out of the bound state channel into dispersive spectral components. The normal form above is the system in which this energy flow is made explicit. Once the normal form (1.15)–(1.16) has been derived, it is natural to seek coupled estimates for $A(t)$ and $\phi_d(t)$ from which their decay can be deduced. This is implemented in section 6. A natural first step is to introduce the auxiliary function

$$\tilde{A}(t) \;\equiv\; e^{\int_0^t\ \rho(s)\ ds}\ A(t) \tag{1.20}$$

for then $\tilde{A}(t)$ satisfies simplified equation of the form

$$\partial_t\tilde{A}(t) \;=\; -\varepsilon^2\Gamma\ \tilde{A}(t) \;+\; \tilde{E}(t;\tilde{A}(t),\phi_d(t)). \tag{1.21}$$

If $\Re\int_0^t\ \rho(s)\ ds$ is uniformly bounded, then modulo time-decay estimates on $\tilde{E}(t;\tilde{A},\phi_d)$ and $F(t;\tilde{A},\phi_d)$, the decay of $\tilde{A}(t)$ and therefore of $A(t)$ follows. For the class of perturbations considered in [22], $\rho(t)$ is a periodic function, having only a finite number of commensurate frequencies, none of them zero. Therefore, in this case $\Re\int_0^t\ \rho(s)\ ds$ is uniformly bounded. However, in the present case $\rho(t)$ is almost periodic with mean $M(\Re\rho)=0$ (see section 9); $\rho(t)$ is displayed in (4.12). $\Re\rho(t)$ has, in general, infinitely many frequencies, $\mu_k-\mu_j,\ k\neq j$ which may accumulate at zero. Most delicate is the case where, along some subsequence, $\mu_k-\mu_j\to 0$. It is well known that the integral of an almost periodic function of mean zero is not necessarily bounded [2], so we are in need of a strategy for estimating the effects of $\Re \int_0^t \rho(s)\, ds$. We address the estimation of $\Re \int_0^t \rho(s)\, ds$ in two different ways corresponding to Theorem 2.3 (see also section 5.1) and Theorems 2.1–2.2 (see also section 5.2). In section 5.1 $\Re \int_0^t \rho(s)\, ds$ is estimated under the hypothesis **(H6)** which requires that the rate of accumulation of a subset of frequencies $\{\mu_j\}_{j\in I}$ is balanced by the decay of the Fourier coefficients g_j as $j \to \infty,\ j \in I$. This leads to a bound on $\Re \int_0^t \rho(s)\, ds$ (Proposition 5.2). In section 5.2 the estimates are based on a more refined analysis; the almost periodic function $\rho(t)$ is decomposed into a part with bounded integral and a part which has mean zero. The latter is controlled using results on the rate at which an almost periodic function approaches its mean.

1.4. Fermi golden rule and obstructions to Poincaré continuation. In the theory of ordinary differential equations it is a standard procedure, given a periodic solution of an unperturbed problem, to seek a periodic or almost periodic solution of a slightly perturbed dynamical system. We now investigate this procedure in the context of (1.7) and its solution $b_*(t)$ for $\varepsilon = 0$. Seek a solution of the form

$$\phi(t) \;=\; b_*(t) \;+\; \phi_1(t) \;+\; \mathcal{O}(\varepsilon^2\beta^2). \tag{1.22}$$

Here, $\phi_1 \;=\; \mathcal{O}(\varepsilon\beta)$.[4] Substitution of (1.22) into (1.7) yields the equation

$$i\partial_t\phi_1 \;=\; H_0\phi_1 \;+\; \varepsilon\beta\, g(t)\, b_*(t). \tag{1.23}$$

This equation has a solution in the class of almost periodic solutions of t with values in the Hilbert space $\mathcal{H}$ only if $\beta\, g(t)\, b_*(t)$ is “orthogonal” to the null space of $i\partial_t - H_0$.

We now derive this condition. Let $e(\zeta)$ be a solution of $H_0 e(\zeta) = \zeta e(\zeta)$. Then, taking the scalar product of (1.23) with $e^{-i\zeta t}\, e(\zeta)$ and applying the operator $\lim_{T\uparrow\infty} T^{-1}\int_0^T \cdot\, dt$ to the resulting equation gives

$$0 \;=\; \lim_{T\uparrow\infty}\, T^{-1}\int_0^T\; e^{i\zeta t}\; e^{-i\lambda_0 t}\; g(t)\; dt\;\; (e(\zeta), \beta\psi_0)\,. \tag{1.24}$$

Substitution of the expansion for $g(t)$ yields

$$\sum_{j\in\mathbb{Z}} g_j\; \delta(\zeta, \lambda_0 + \mu_j)\;\; (e(\zeta), \beta\psi_0) \;=\; 0, \tag{1.25}$$

where $\delta(a,b) \;=\; 0$ if $a \;\neq\; b$ and $\delta(a,a) \;=\; 1$. If ζ, which lies in the spectrum of H_0, satisfies $\zeta \;=\; \lambda_0 + \mu_k$ for some $k \in \mathbb{Z}$ (which will be the case in our example if $\lambda_0 + \mu_k > 0$), then we have that

$$(e(\lambda_0 + \mu_k), \beta\psi_0) \;=\; 0 \tag{1.26}$$

is a necessary condition for the existence of a family of solutions of (1.7) which converges to $b_*(t)$ as the perturbation $W(t)$ tends to zero. We immediately recognize the inner product in (1.26) as the projection of $\beta\psi_0$ onto the generalized eigenmode at the resonant frequency $\lambda_0 + \mu_k$ which arises in (1.17); see also (1.18). Therefore the obstruction to continuation of $b_*(t)$ to a nearby almost periodic state of the system can be identified with the damping mechanism.

[4] This argument is heuristic so we do not specify the norm with which the size of β is measured.

1.5. Outline. The paper is structured as follows. In section 2 we give a general formulation of the problem. The hypotheses on H_0, the unperturbed Hamiltonian, and $W(t)$, the perturbation, are introduced and discussed. There are two types of theorems: Theorems 2.1 and 2.2 and Theorem 2.3. Although the conclusions of these are quite similar, as discussed above, they differ in a key hypothesis on the perturbation $W(t)$, which is relevant in the case where $W(t)$ has infinitely many frequencies which may resonate with the continuous spectrum. In section 3 we apply the results of section 2 to the case of Schrödinger operators $H_0 = -\Delta + V(x)$ defined on $L^2(\mathbb{R}^3)$. To check the key local energy decay hypotheses we use results of Jensen and Kato [5] on expansions of the resolvent of H_0 near zero energy, the edge of the continuous spectrum. In section 4 the dynamical system (1.3) is reformulated as a system governing the interaction of the bound state and dispersive part of the solution. This section contains an important computation in which the key resonance is made explicit and a perturbed "normal form" for the bound state evolution is derived (Proposition 4.1). Sections 5 and 6 contain estimates for the bound state and dispersive parts of the solution for intermediate and large time scales. In section 7 we discuss extensions of our Theorems 2.1–2.3 to a more general class of perturbations. We shall frequently make use of some singular operators which are rigorously defined in section 8, an appendix, and of elements of the theory of almost periodic functions [2, 9], which are assembled in section 9, the second appendix.

Notations and terminology. Throughout this paper we will use the following notations:

$\mathbb{N} = \{1,\ 2,\ 3,\ldots\}$;

$\mathbb{N}_0 = \{0,\ 1,\ 2,\ 3,\ldots\}$;

$\mathbb{Z} = \{\ldots,-3,\ -2,\ -1,\ 0,\ 1,\ 2,\ 3,\ldots\}$;

for z a complex number, $\Re z$ and $\Im z$ denote, respectively, its real and imaginary parts; a generic constant will be denoted by C, D, etc;

$\langle x\rangle = \left(1+|x|^2\right)^{\frac{1}{2}}$;

$\mathcal{L}(\mathcal{A},\mathcal{B}) =$ the space of bounded linear operators from $\mathcal{A}$ to $\mathcal{B}$; $\mathcal{L}(\mathcal{A},\mathcal{A}) \equiv \mathcal{L}(\mathcal{A})$.

Functions of self-adjoint operators are defined via the spectral theorem; see, for example, [17]. The operators containing boundary value of resolvents or singular distributions applied to self-adjoint operators are defined in section 8.

2. General formulation and main results.

Consider the general system

$$\begin{aligned} i\partial_t\phi(t) &= \left(H_0 + W(t)\right)\phi(t),\\ \phi|_{t=0} &= \phi(0). \end{aligned} \tag{2.1}$$

Here, $\phi(t)$ denotes a function of time, t, with values in a complex Hilbert space $\mathcal{H}$.

Hypotheses on H_0.

(H1) H_0 is self-adjoint on $\mathcal{H}$ and both H_0 and $W(t),\ t\in\mathbb{R}^1$, are densely defined on a subspace $\mathcal{D}$ of $\mathcal{H}$.

The norm on $\mathcal{H}$ is denoted by $\|\cdot\|$ and the inner product of $f,g\in\mathcal{H}$, by (f,g).

(H2) The spectrum of H_0 is assumed to consist of an absolutely continuous part, $\sigma_{\rm cont}(H_0)$, with associated spectral projection $\mathbf{P_c}$ and a single isolated eigenvalue λ_0 with corresponding normalized eigenstate, ψ_0, i.e.,

$$H_0\psi_0 = \lambda_0\psi_0,\ \|\psi_0\| = 1. \tag{2.2}$$

The manner in which we shall measure the decay of solutions is typically in a local decay sense, e.g., for the scalar Schrödinger equation governing a function defined on

$\mathbb{R}^n$ we measure local decay using the norms $f \mapsto \|\langle x\rangle^{-s} f\|_{L^2}$, where $s > 0$. So that our theory applies to a class of general systems (involving, for example, vector equations with matrix operators), we assume the existence of self-adjoint "weights" w_- and w_+ such that

(i) w_+ is defined on a dense subspace of $\mathcal{H}$ and on which $w_+ \geq cI, \ \ c > 0$.

(ii) $w_- \in \mathcal{L}(\mathcal{H})$ such that $\mathrm{Range}(w_-) \subseteq \mathit{Domain}(w_+)$.

(iii) $w_+ \ w_- \ \mathbf{P_c} \ = \ \mathbf{P_c}$ on $\mathcal{H}$ and $\mathbf{P_c} \ = \ \mathbf{P_c} \ w_- \ w_+$ on the domain of w_+.

In the scalar case, w_+ and w_- correspond to multiplication by $\langle x\rangle^s$ and $\langle x\rangle^{-s}$, respectively; see section 3.

The following hypothesis ensures that the unperturbed dynamics satisfies sufficiently strong dispersive time-decay estimates. Let $\{\mu_j\}_{j\in\mathbb{Z}}$ denote the set of Fourier exponents associated with the perturbation W (see hypothesis **(H4)** below).

(H3) *Local decay estimates on* e^{-iH_0t}.

Let $r_1 > 1$. There exist w_+ and w_-, as above, and a constant $\mathcal{C}$ such that for all $f \in \mathcal{H}$ satisfying $w_+ f \in \mathcal{H}$ we have

$$\text{(2.3)} \quad (\mathbf{a}) \quad \|w_- e^{-iH_0t}\mathbf{P_c} f\| \leq \mathcal{C}\ \langle t\rangle^{-r_1}\|w_+ f\| \text{ for } t \in \mathbb{R};$$

$$\text{(2.4)} \quad (\mathbf{b}) \quad \|w_- e^{-iH_0t}(H_0 - \lambda_0 - \mu_j - i0)^{-1}\ \mathbf{P_c} f\| \leq\ \mathcal{C}\ \langle t\rangle^{-r_1}\ \|w_+ f\| \text{ for } t \geq 0$$

and for all $j \in \mathbb{Z}$. For $t < 0$ estimate (2.4) is assumed to hold with $-i0$ replaced by $+i0$. See section 8 for the definition of the singular operator in (2.4).

Remark 2.1. There is a good deal of literature on local energy decay estimates of the form (2.3) for $e^{-iH_0t}\mathbf{P_c}$ in the case $H_0 = \ -\Delta + V(x)$ on $L^2(\mathbb{R}^n)$. These results require sufficient regularity and decay of the potential $V(x)$. We refer the reader to [5, 6] and [14]; see also [16, 18].

Remark 2.2. Estimates of the type **(H3b)** are obtained in [22], [23, Appendix A]. A key point here is that we require that one can choose the constant, $\mathcal{C}$, in (2.4) to hold for all μ_j. It appears difficult to deduce this uniformity of the constant by the general arguments used in [22] and [23]. However, in section 3, where we apply our results to a class of Schrödinger operators, we can verify (**H3b**) using known results on the spectral measure.

(H4) Hypotheses on the perturbation $W(t)$.

We consider time-dependent symmetric perturbations of the form

$$\text{(2.5)} \qquad W(t) = \frac{1}{2}\beta_0 + \sum_{j\in\mathbb{N}} \cos(\mu_j t)\ \beta_j \text{ with } \beta_j^* = \beta_j \text{ and } \sum_{j\in\mathbb{N}_0} \|\beta_j\|_{\mathcal{L}(\mathcal{H})} < \infty.$$

In many applications, β_j are spatially localized scalar or matrix functions. Note that formula (2.5) can be rewritten in the form

$$\text{(2.6)} \qquad W(t) = \frac{1}{2}\ \sum_{j\in\mathbb{Z}} \exp(-i\mu_j t)\beta_j,$$

where $\mu_0 = 0$ and for $j < 0$, $\mu_j = -\mu_{-j}, \ \beta_j = \beta_{-j}$. Thus, $W(t)$ is an almost periodic function with values in the Banach space $\mathcal{L}(\mathcal{H})$ with the Fourier exponents $\{\mu_j\}_{j\in\mathbb{Z}}$ and corresponding Fourier coefficients $\{\beta_j\}_{j\in\mathbb{Z}}$; see, for example, [9].

To measure the size of the perturbation W, we introduce the norm

$$\text{(2.7)} \qquad |||W|||\ \equiv\ \frac{1}{2}\sum_{j\in\mathbb{Z}} \|w_+\ \beta_j\|_{\mathcal{L}(\mathcal{H})}\ +\ \frac{1}{2}\sum_{j\in\mathbb{Z}} \|\ \beta_j\ \|_{\mathcal{L}(\mathcal{H}_-,\mathcal{H}_+)},$$

which is assumed to be finite. Here $\mathcal{H}_+$, respectively, $\mathcal{H}_-$, denote the closure of the domain of w_+, respectively, the range of $\mathbf{P_c}$, with norm $f \to \|w_+ f\|$, respectively, $f \to \|w_- f\|$.

Remark 2.3. A special case which arises in various models is

$$W(t) \;=\; g(t)\beta, \tag{2.8}$$

where

$$g(t) \;=\; \sum_j \, g_j \cos \mu_j t, \tag{2.9}$$

$\|w_+\beta\|_{\mathcal{L}(\mathcal{H})} + \;\|\beta\|_{\mathcal{L}(\mathcal{H}_-,\mathcal{H}_+)} \;<\; \infty$ and the sequence $\{g_j\}$ is absolutely summable.

Remark 2.4. Our results are valid in the more general case

$$W(t) = \frac{1}{2}\beta_0 + \sum_{j\in\mathbb{N}} \cos(\mu_j t + \delta_j)\;\beta_j,$$

where β_j are self-adjoint such that expression (2.7) is finite. This follows because the proofs use only the self-adjointness of W and the expansion

$$W(t) = \frac{1}{2}\,\sum_{j\in\mathbb{Z}} \exp(-i\mu_j t)\tilde{\beta}_j,$$

where $\tilde{\beta}_j = e^{-i\mathrm{sgn}(j)\delta_j}\beta_j$ and $\mu_{-j} = -\mu_j,\; \mu_0 = 0$.

We will impose a resonance condition which says that $\{\lambda_0 + \mu_j\}_{j\in\mathbb{Z}} \cap \sigma_{\text{cont}}(H_0)$ is nonempty and that there is nontrivial coupling; see section 1.4. Let us first denote by I_{res} the following set:

$$I_{res} \;=\; \{j \in \mathbb{Z} \;:\; \lambda_0 + \mu_j \in \sigma_{\text{cont}}(H_0)\}. \tag{2.10}$$

(H5) *Resonance condition.* Fermi golden rule.

I_{res} is nonempty and furthermore, there exists $\theta_0 > 0$, independent of W, such that

$$\Gamma \;\equiv\; \frac{\pi}{4} \sum_{j\in I_{res}} (\mathbf{P_c}\beta_j\psi_0,\; \delta(H_0 - \lambda_0 - \mu_j)\mathbf{P_c}\beta_j\psi_0.) \geq \theta_0 |||W|||^2 > 0. \tag{2.11}$$

Remark 2.5. For the exact definition of the Dirac-type operator in (2.11), see section 8. That Γ is finite is a consequence of the estimate (8.8) and

$$\Gamma \leq \frac{C_0}{\pi}\sum_j \|w_+\beta_j\|^2 \leq \frac{C_0}{\pi}|||W|||^2; \tag{2.12}$$

see also [1].

We now state our main results.

THEOREM 2.1. *Let us fix H_0 and $W(t)$ satisfying hypotheses* **(H1)–(H5)**. *Consider the initial value problem*

$$\begin{aligned} &i\partial_t\phi(t) = (H_0 + \varepsilon W(t))\,\phi(t),\\ &\phi|_{t=0} = \phi(0) \end{aligned} \tag{2.13}$$

with $w_+\phi(0) \in \mathcal{H}$. Then, there exists an $\varepsilon_0 > 0$ (depending on $\mathcal{C}$, r_1, and θ_0) such that whenever $|\varepsilon| < \varepsilon_0$, the solution, $\phi(t)$, of (2.13) *satisfies the local decay estimate*

$$\|w_- \ \phi(t)\| \le C\langle t\rangle^{-r_1}\|w_+ \ \phi(0)\|, \quad t \in \mathbb{R}. \tag{2.14}$$

Under the same hypotheses as Theorem 2.1, we obtain more detailed information on the behavior of $\phi(t)$.

THEOREM 2.2. *Assume the hypotheses of Theorem* 2.1. *For any $0 < \gamma < \Gamma$ there exist the constants C and D (depending on $\mathcal{C}$, r_1, θ_0, and γ) such that any solution of* (2.13)*, for $|\varepsilon| < \varepsilon_0$ and $w_+\phi(0) \in \mathcal{H}$, satisfies*

$$\begin{aligned} \phi(x,t) &= a(t)\psi_0 + \phi_d(t), \quad (\psi_0\ ,\ \phi_d(t),\,) \ = \ 0, \\ a(t) &= \ a(0)\ e^{-\varepsilon^2(\Gamma-\gamma)|t|}e^{i\omega(t)} \ + \ R_a(t), \\ P(t) &= \ P(0)\ e^{-2\varepsilon^2(\Gamma-\gamma)|t|} \ + \ R'_a(t), \\ \phi_d(t) &= e^{-iH_0t}\ \mathbf{P_c}\phi(0) \ + \ \tilde{\phi}(t), \end{aligned} \tag{2.15}$$

where Γ is given by (2.11) *and $\omega(t)$ is a real-valued phase given by*

$$\begin{aligned} \omega(t) \ &= \ -\lambda_0 t - \varepsilon\left(\psi_0,\ \int_0^t W(s)ds\ \psi_0\right) \\ &+\frac{1}{4}\varepsilon^2 t\sum_{j\in\mathbb{Z}}\left(\beta_j\psi_0, \text{P.V.}(H_0-\lambda_0-\mu_j)^{-1}\mathbf{P_c}\beta_j\psi_0\right) \\ &+\ \frac{1}{4}\varepsilon^2\ \Re\int_0^t \sum_{j,k\in\mathbb{Z},j\neq k} e^{i(\mu_k-\mu_j)t}\left(\beta_k\psi_0,(H_0-\lambda_0-\mu_j-i0)^{-1}\mathbf{P_c}\beta_j\psi_0\right). \end{aligned} \tag{2.16}$$

$P(t)$ is defined in (1.4) *and for any fixed $T_0 > 0$ we have*

$$|R_a(t)| \ \le \ C\ |\varepsilon|\ |||W|||,\ |t| \le \frac{T_0}{\varepsilon^2\Gamma}, \tag{2.17}$$

$$|R'_a(t)| \ \le \ D\ |\varepsilon|\ |||W|||,\ |t| \le \frac{T_0}{\varepsilon^2\Gamma}. \tag{2.18}$$

Moreover,

$$|R_a(t)| = \mathcal{O}(\langle t\rangle^{-r_1}),\ |R'_a(t)| = \mathcal{O}(\langle t\rangle^{-r_1}),\ |t| \to \infty.$$

Finally, $\tilde{\phi} = \phi_1 + \phi_2$ is given in (4.9) *with $\|w_-\tilde{\phi}(t)\| = \mathcal{O}(\langle t\rangle^{-r_1})$ as $|t| \to \infty$. Therefore, by* **(H3)** *$\|w_-\phi_d(t)\| = \mathcal{O}(\langle t\rangle^{-r_1})$ as $|t| \to \infty$.*

Remark 2.6. Suppose the initial data is given by the bound state of the unperturbed problem, i.e., $\phi(x,0) = \psi_0(x)$, $a(0) = 1$, $\phi_d(0) = 0$. Then, from the expansion of the solution we have that for $0 \le t \le \varepsilon^{-2}\Gamma^{-1}$ that $P(t)$ (see (1.4)) is of order $e^{-2\varepsilon^2(\Gamma-\gamma)t}$ with an error of order ε. Hence it is natural to view the state $\psi_0 e^{-i\lambda_0 t}$ as a metastable state with lifetime $\tau = \varepsilon^{-2}(\Gamma-\gamma)^{-1} \sim \varepsilon^{-2}|||W|||^{-2}$. Although $\gamma > 0$ is arbitrary we have not inferred that the actual lifetime is $\tau = \varepsilon^{-2}\Gamma^{-1}$ under hypotheses **(H1)**–**(H5)**. The reason is that the constants C and D in the estimates (2.17) and (2.18) blow up as $\gamma \searrow 0$. In order to remedy this we need an additional hypothesis.

(H6) *Control of small denominators.*

There exists $\xi > 0$, independent of W, such that

$$\sum_{j\in I_{res},\ k\in\mathbb{Z},\ j\neq k}\left|\frac{1}{\mu_j-\mu_k}(\mathbf{P_c}\beta_k\psi_0, \delta(H_0-\lambda_0-\mu_j)\mathbf{P_c}\beta_j\psi_0)\right| \le \xi\ |||W|||^2. \tag{2.19}$$

Remark 2.7. By (8.8) we have that

$$\sum_{j\in I_{res},\ k\in\mathbb{Z},\ j\neq k} |(\beta_k\psi_0, \delta(H_0-\lambda_0-\mu_j)\beta_j\psi_0)| \leq C\ \pi^{-1}\ |||W|||^2 \tag{2.20}$$

is finite (see also Remark 2.5). Thus, **(H6)** is important only if

$$\inf\{|\mu_j-\mu_k| :\ j,k\in\mathbb{Z},\ j\neq k \text{ and } \lambda_0+\mu_j\in\sigma_{\text{cont}}(H_0)\} = 0, \tag{2.21}$$

i.e., the Fourier exponents $\{\mu_j\}$ are such that $\lambda_0+\mu_j$ accumulate in σ_c. In particular, if the perturbation $W(t)$ consists of a trigonometric polynomial

$$W(t)\ =\ \sum_{j=1}^{N} \cos\mu_j t\ \beta_j, \tag{2.22}$$

then **(H6)** is trivially satisfied.

Remark 2.8. Hypothesis **(H6)** can be imposed by balancing the clustering of the frequencies $\lambda_0+\mu_j$ in the continuous spectrum of H_0 with rapid decay of $(\beta_k\psi_0\ ,\ \delta(H_0-\lambda_0-\mu_j)\ \beta_j\psi_0)$ as $j,k\to\infty$. Let $\beta_j(x) = g_j\ \beta(x)$. Then, $W(t,x) = \sum_j\ g_j\cos(\mu_j t)\ \beta(x)$. Using Remark 2.5 we find that the left-hand side of (2.19) is bounded by $\sum_{j,k\in\mathbb{Z};j\neq k}|g_jg_k|\ |\mu_j-\mu_k|^{-1}\ |||W|||^2$. The constant ξ in (2.19) is finite if, for example, $\mu_j = 2|\lambda_0|+|j|^{-1},\ g_j = |j|^{-2-\tau},\ \tau>0$.

In case **(H6)** is satisfied we have the following improvement of Theorem 2.2.

Theorem 2.3. *Assume the hypotheses* **(H1)**–**(H6)** *hold. Then there exist* ε_0 *and the constants* C, D *(depending on* $\mathcal{C}$, r_1, θ_0 *and* ξ*) such that any solution of* (2.13)*, for* $|\varepsilon|<\varepsilon_0$ *and* $w_+\phi(0)\in\mathcal{H}$*, satisfies*

$$\begin{aligned}
\phi(x,t) &= a(t)\psi_0+\phi_d(t),\quad (\psi_0\ ,\ \phi_d(t))\ =\ 0,\\
a(t) &= a(0)\ e^{-\varepsilon^2\Gamma|t|}\ e^{i\omega(t)}\ +\ R_a(t),\\
P(t) &= P(0)\ e^{-2\varepsilon^2\Gamma|t|}\ +\ R'_a(t),\\
\phi_d(t) &= e^{-iH_0t}\ \mathbf{P_c}\phi(0)\ +\ \tilde{\phi}(t).
\end{aligned} \tag{2.23}$$

Here, $\omega(t)$ *is given by* (2.16) *and* $R_a(t)$, $R'_a(t)$, $w_-\phi_d(t)$ *satisfy the estimates of Theorem* 2.2.

3. An application: The Schrödinger equation. In this section we verify hypotheses **(H1)**–**(H4)** in the particular case of the Schrödinger equation on the three-dimensional space with a time almost periodic and spatially localized perturbing potential:

$$i\partial_t\phi\ =\ (-\Delta+V(x))\,\phi\ +\ \varepsilon W(x,t)\phi \tag{3.1}$$

with $\phi:\mathbb{R}^3\times\mathbb{R}\to\mathbb{C},\ (x,t)\to\phi(x,t)$, and

$$W(x,t) = \frac{1}{2}\beta_0(x)+\sum_{j\in\mathbb{N}}\cos(\mu_jt)\beta_j(x),$$

where $\mu_j\in\mathbb{R},\ j\in\mathbb{N}_0$, and $\beta_j:\mathbb{R}^3\to\mathbb{R},\ j\in\mathbb{N}$, are localized functions. Models of the sort considered in this example occur in the study of ionization of an atom by a time-varying electric field; see [10, 4].

We take $\mathcal{H} = L^2(\mathbb{R}^3)$ and $H_0 \equiv -\Delta + V(x)$, where $V(x)$ is real-valued with moderately short range. More precisely, we suppose that there exists $\sigma > 4$ and a constant D such that

$$|V(x)| \le D(1+|x|)^{-\sigma}. \tag{3.2}$$

Thus, H_0 is self-adjoint and densely defined in L^2. In what follows we assume that H_0 has exactly one eigenvalue which is strictly negative and that the remainder of the spectrum is absolutely continuous and equal to the positive half-line. Our results can be extended to operators with strictly negative, multiple eigenvalues [7].

We first discuss the local decay hypothesis **(H3)**. As weights used to measure local energy decay we take $w_\pm \equiv \langle x\rangle^{\pm s}$, where $s > 7/2$ and fix $r_1 = 3/2$. Our aim is to obtain the estimates

$$\textbf{(H3a)}\quad \|w_- e^{-iH_0t}\mathbf{P_c} f\| \le \mathcal{C}\ \langle t\rangle^{-3/2}\|w_+ f\|, \tag{3.3}$$

$$\textbf{(H3b)}\quad \|w_- e^{-iH_0t}(H_0-\lambda_0-\mu_j-i0)^{-1}\ \mathbf{P_c} f\| \le\ \mathcal{C}\ \langle t\rangle^{-3/2}\ \|w_+ f\| \tag{3.4}$$

for all $\mu_j \in \mathbb{Z}$ with $\mathcal{C}$ independent of j.

We shall assume that the frequencies $\{\lambda_0 + \mu_j\}$ do not accumulate at zero, the edge of the continuous spectrum of H_0:

$$m_* \ \equiv\ \min\{\ |\lambda_0+\mu_j|\ :\ j\in\mathbb{Z}\ \} >\ 0. \tag{3.5}$$

To prove (3.3) and (3.4) we use the spectral representation for the operators $e^{-iH_0t}\ \mathbf{P_c}$ and $e^{-iH_0t}\ (H_0-\lambda_0-\mu_j-i0)^{-1}\mathbf{P_c}$, namely,

$$e^{-iH_0t}\mathbf{P_c} = \int_0^\infty e^{-i\lambda t}E'(\lambda)d\lambda, \tag{3.6}$$

$$e^{-iH_0t}(H_0-\lambda_0-\mu_j-i0)^{-1}\mathbf{P_c} = \int_0^\infty e^{-i\lambda t}(\lambda-\lambda_0-\mu_j-i0)^{-1}E'(\lambda)d\lambda, \tag{3.7}$$

where $E'(\lambda) = \pi^{-1}\ \Im(H_0-\lambda-i0)^{-1}$ is the spectral density induced by H_0, [5].

The technique of getting **(H3a)** from (3.6) is presented in [5, section 10] and it can be summarized in the following way. We decompose the integral in (3.6) in two parts, corresponding to low energies (λ near zero) and high energies (λ away from zero) by writing

$$\begin{aligned} E' &= \chi E' + (1-\chi)E', \\ w_- e^{-iH_0t}\mathbf{P_c} w_- &= \int_0^\infty e^{-i\lambda t}\chi(\lambda) w_- E'(\lambda) w_- d\lambda + \int_0^\infty e^{-i\lambda t}(1-\chi(\lambda)) w_- E'(\lambda) w_- d\lambda \\ &=\ S_1\ +\ S_2. \end{aligned} \tag{3.8}$$

Here, $\chi(\lambda)$ is a smoothed characteristic function of a neighborhood of origin, chosen so that

$$\chi(\lambda)\ \equiv\ 1,\ |\lambda|\ \le\ \frac{1}{2}m_*,$$
$$\chi(\lambda)\ \equiv\ 0,\ |\lambda|\ \ge\ \frac{3}{4}m_*.$$

To estimate the two integrals in (3.8) we make use of the detailed results of [5] on the family of operators $\{E'(\lambda)\}$. First, by Theorem 8.1 and Corollary 8.2 of [5], $w_-\partial_\lambda^k E'(\lambda)w_-$ is bounded on L^2 and satisfies

$$\|\ w_-\partial_\lambda^k E'(\lambda)w_-\ \|_{\mathcal{B}(L^2)} = O(\lambda^{-(k+1)/2}) \text{ as } \lambda\to\infty \tag{3.9}$$

for $k \in \{0,1,2,3\}$. Integration by parts twice in the second integral in (3.8) and use of the estimate (3.9) with $k=2$ yields the estimate

$$\| S_2 \|_{\mathcal{B}(L^2)} \;=\; o(t^{-2}) \text{ as } t \to \infty. \tag{3.10}$$

Next, by Theorem 6.3 of [5] we have the low energy asymptotic expansion

$$w_- E'(\lambda) w_- = -\lambda^{-1/2} B_{-1} + \lambda^{1/2} B_1 + o(\lambda^{1/2}) \text{ as } \lambda \to 0, \tag{3.11}$$

where $B_{-1},\ B_1$ are bounded linear operators on L^2. Use of this expansion in the first integral of (3.8) yields the expansion in $\mathcal{B}(L^2)$:

$$S_1 \;=\; (\pi i)^{-1/2} t^{-1/2} B_{-1} - (4\pi i)^{-1/2} t^{-3/2} B_1 + o(t^{-3/2}) \text{ as } t \to \infty. \tag{3.12}$$

Thus, **(H3a)** is satisfied provided that B_{-1} is the null operator or equivalently $H_0\psi = 0$ has no solution with the property $w_-\psi \in L^2(\mathbb{R}^3)$. The last condition holds for generic potentials $V(x)$ and when it is violated one says that H_0 has *zero energy resonance*; see [5] for details.

In the same way one can prove **(H3b)** from the spectral representation (3.7) provided that the integral is nonsingular, i.e., $\lambda_0+\mu_j < 0$. In the case $\lambda_0+\mu_j \geq m_* > 0$ we first decompose the singular integral in two parts, one away from singularity point, $\lambda_0+\mu_j$, and the other in a neighborhood of it by using the smoothed characteristic function

$$\chi_j(\lambda) \;=\; \chi(\lambda - \lambda_0 - \mu_j), \tag{3.13}$$

which is supported in a neighborhood of $\lambda_0+\mu_j$, which does not include $\lambda = 0$:

$$\begin{aligned} e^{-iH_0 t}(H_0 - \lambda_0 - \mu_j - i0)^{-1}\mathbf{P_c} = \int_0^\infty & e^{-i\lambda t}(\lambda - \lambda_0 - \mu_j)^{-1}(1-\chi_j(\lambda))E'(\lambda)d\lambda \\ + \int_0^\infty & e^{-i\lambda t}(\lambda - \lambda_0 - \mu_j - i0)^{-1}\chi_j(\lambda)E'(\lambda)d\lambda. \end{aligned} \tag{3.14}$$

The nonsingular integral may be treated as above while the singular one defines the singular operator

$$T_j = e^{-iH_0 t}(H_0 - \lambda_0 - \mu_j - i0)^{-1}\chi_j(H_0)\mathbf{P_c}$$

via the spectral theorem. Here, $T_j = \lim_{\eta\searrow 0} T_j^\eta$, where

$$T_j^\eta \equiv e^{-iH_0 t}(H_0 - \lambda_0 - \mu_j - i\eta)^{-1}\chi_j(H_0)\mathbf{P_c}.$$

To estimate its L^2 operator norm we use the integral representation

$$w_- T_j^\eta w_- = \frac{1}{i}\int_t^\infty e^{i(\lambda_0+\mu_j+i\eta)(s-t)} w_- e^{-iH_0 s}\chi_j(H_0)\mathbf{P_c} w_- ds. \tag{3.15}$$

But this reduces to the evaluation of

$$w_- e^{-iH_0 s}\chi_j(H_0)\mathbf{P_c} w_- = \int_0^\infty e^{-i\lambda s}\chi_j(\lambda) w_- E'(\lambda) w_- d\lambda,\ s \geq t, \tag{3.16}$$

where we used again the spectral representation theorem. Integration by parts three times in (3.16) and use of the estimate (3.9) with $k=3$ implies

$$\|w_- e^{-iH_0 s}\chi_j(H_0)\mathbf{P_c} w_-\|_{\mathcal{B}(L^2)} = o(s^{-3}) \text{ as } t \to \infty.$$

Replacing this in (3.15), integrating and passing to the limit as $\eta \searrow 0$ we obtain an $o(t^{-2})$ estimate for T_j which is even better than we need to satisfy **(H3b)**.

Moving now towards hypothesis **(H4)**, we may choose the time-dependent perturbation to be of the form

$$W(x,t) \;=\frac{1}{2}\beta_0 \;+\; \sum_{j\in\mathbb{N}}\cos\mu_j t\;\beta_j(x) \tag{3.17}$$

with β_j rapidly decaying in x, e.g., $\langle x\rangle^{2s}\|\beta_j(x)\| \le C_j$ for all $x\in\mathbb{R}^3$, $j\in\mathbb{N}_0$, where $\sum_{j\in\mathbb{N}_0} C_j < \infty$. Thus, **(H4)** is satisfied as well.

Therefore, our main results Theorems 2.1–2.2 on the structural instability of the unperturbed bound state and large time behavior for systems of the form (3.1) apply provided **(H5)**, the Fermi golden rule resonance condition, holds. For results concerning more general perturbations than the ones in (3.1) see section 7.

4. Decomposition and derivation of the dispersive normal form. The results of this section rely on hypothesis **(H1)** through **(H4)** only, so they may and will be used in proving Theorems 2.1–2.3.

As in [21, 22] and [23], we begin by deriving a decomposition of the solution, $\phi(t)$, which will facilitate the study of its large time behavior. Let

$$\phi(t) = a(t)\psi_0 + \phi_d(t) \tag{4.1}$$

with the orthogonality condition

$$(\psi_0, \phi_d(t)) = 0 \;\text{ for all }\; t. \tag{4.2}$$

Note therefore that $\phi_d \;=\; \mathbf{P_c}\phi_d$.

We proceed by first inserting (4.1) into (2.13), which yields the equation

$$\begin{aligned} i\partial_t a(t)\psi_0 \;+\; i\partial_t\phi_d(t) \;&=\; \lambda_0 a(t)\psi_0 \;+ H_0\phi_d(t)\\ &\quad+\varepsilon a(t)W(t)\psi_0 \;+\; \varepsilon W(t)\phi_d(t). \end{aligned} \tag{4.3}$$

Taking the inner product of (4.3) with ψ_0 we get the following equation for $a(t)$:

$$\begin{aligned} i\partial_t a \;&=\; \lambda_0 a(t) \;+\; \varepsilon\,(\psi_0, W(t)\psi_0)\,a(t) \;+\; \varepsilon\,(\psi_0, W(t)\phi_d)\,,\\ a(0) \;&=\; (\psi_0, \phi(0))\,. \end{aligned} \tag{4.4}$$

In deriving (4.4) we have used that ψ_0 is normalized and the relation

$$(\psi_0, \partial_t\phi_d) = 0, \tag{4.5}$$

a consequence of (4.2).

Applying $\mathbf{P_c}$ to (4.3), we obtain an equation for ϕ_d:

$$\begin{aligned} i\partial_t\phi_d(t) \;&=\; H_0\phi_d(t) \;+\; \varepsilon\mathbf{P_c}W(t)\phi_d(t) \;+\; \varepsilon a(t)\mathbf{P_c}W(t)\psi_0,\\ \phi_d(0) \;&=\; \mathbf{P_c}\phi(0). \end{aligned} \tag{4.6}$$

Since we are after a slow resonant decay phenomenon, it will prove advantageous to extract the fast oscillatory behavior of $a(t)$. We therefore define

$$A(t) \equiv e^{i\lambda_0 t}a(t). \tag{4.7}$$

Then, (4.4) reads

$$\partial_t A \; = \; -i\varepsilon A\left(\psi_0, W(t)\psi_0\right) \; - \; i\varepsilon e^{i\lambda_0 t} \; \left(\psi_0, W(t)\phi_d(t)\right). \tag{4.8}$$

Solving (4.6) by Duhamel's formula we have

$$\begin{aligned}\phi_d(t) = e^{-iH_0t}\phi_d(0) \; - \; i\varepsilon \int_0^t e^{-iH_0(t-s)}\mathbf{P_c}W(s)a(s)\psi_0 ds \\ -i\varepsilon \int_0^t e^{-iH_0(t-s)}\mathbf{P_c}W(s)\phi_d(s) \; ds \\ \equiv \phi_0(t) \; + \; \phi_1(t) \; + \; \phi_2(t).\end{aligned} \tag{4.9}$$

By standard methods, the system (4.8)–(4.9) for $A(t)$ and $\phi_d(t) = \phi(t) - e^{-i\lambda_0 t} \, A(t) \, \psi_0$ has a global solution in t with

$$A \in C^1(\mathbb{R}), \; \|\phi_d(t)\| \in C^0(\mathbb{R}), \; \|w_-\phi_d(t)\| \in C^0(\mathbb{R}).$$

Our analysis of the $|t| \to \infty$ behavior is based on a study of this system.

By inserting (4.9) into (4.8) we get

$$\partial_t A(t) \; = \; -i\varepsilon A(t)\left(\psi_0, W(t)\psi_0\right) \; - \; i\varepsilon e^{i\lambda_0 t}\sum_{j=0}^{2}\left(\psi_0, W(t)\phi_j\right). \tag{4.10}$$

We next give a detailed expansion of the sum in (4.10). It is in the $j = 1$ term that the key resonance is found. This makes it possible to find a normal form for (4.10) in which *internal damping* in the system is made explicit. This damping reflects the transfer of energy from the discrete to continuum modes of the system and the associated radiative decay of solutions.

PROPOSITION 4.1. *For $t > 0$,*

$$\partial_t A(t) \; = \; \left(-\varepsilon^2\Gamma \; + \; \rho(t)\right) \; A(t) \; + \; E(t), \tag{4.11}$$

where Γ is defined in (2.11),

$$\begin{aligned}\rho(t) \; &= \; -i\varepsilon\left(\psi_0, \; W(t)\psi_0\right) \\ &+ \; \frac{i}{4}\varepsilon^2\sum_{j\in\mathbb{Z}}\left(\beta_j\psi_0, \text{P.V.}(H_0 - \lambda_0 - \mu_j)^{-1}\mathbf{P_c}\beta_j\psi_0\right) \\ &+ \; \frac{i}{4}\varepsilon^2\sum_{j,k\in\mathbb{Z}, j\neq k} e^{i(\mu_k - \mu_j)t}\left(\beta_k\psi_0, (H_0 - \lambda_0 - \mu_j - i0)^{-1}\mathbf{P_c}\beta_j\psi_0\right)\end{aligned} \tag{4.12}$$

and

$$\begin{aligned}E(t) \; = \; &-\frac{i}{4}\varepsilon^2 A(0)e^{i\lambda_0 t}\sum_{j,k\in\mathbb{Z}} e^{i\mu_k t} \; \left(\beta_k\psi_0, \; e^{-iH_0 t} \; (H_0 - \lambda_0 - \mu_j - i0)^{-1}\mathbf{P_c}\beta_j\psi_0\right) \\ &-\frac{i}{4}\varepsilon^2 e^{i\lambda_0 t}\sum_{j,k\in\mathbb{Z}} e^{i\mu_k t} \\ &\left(\beta_k\psi_0, \int_0^t e^{-iH_0(t-s)}(H_0 - \lambda_0 - \mu_j - i0)^{-1}\mathbf{P_c}e^{-i(\lambda_0+\mu_j)s}\partial_s A(s)\beta_j\psi_0\right) ds \\ &-i\varepsilon e^{i\lambda_0 t} \; \left(\psi_0, \; W(t)\phi_0(t)\right) \\ &-i\varepsilon e^{i\lambda_0 t} \; \left(\psi_0, \; W(t)\phi_2(t)\right).\end{aligned} \tag{4.13}$$

Here, ϕ_0 and ϕ_2 are given in (4.9).

Although the proposition is stated for $t > 0$, an analogous proposition with $-\varepsilon^2\Gamma$ replaced by $\varepsilon^2\Gamma$ holds for $t < 0$. The modification required to treat $t < 0$ is indicated in the proof.

Remark 4.1. (1) The point of (4.11) is that the source of damping, $\Gamma > 0$, which arises due to the coupling of the discrete bound state to the continuum modes by the almost periodic perturbation is made explicit. Note that $\Re\rho(t)$ is of order $\varepsilon^2|||W|||^2$ as the first two terms of $\rho(t)$ are purely imaginary inducing only a phase shift in the solution, $A(t)$. The last term is of the same order as the damping and may compete with it. A key point of our analysis is to assess the contribution of this last term in (4.12).

(2) The leading order part of (4.11) is the analogue of the dispersive normal form derived in [24] for a class of nonlinear dispersive wave equations.

Proof of Proposition 4.1. Using the expression for $W(t)$ in (2.6), which is a uniform convergent series with respect to $t \in \mathbb{R}$, and the definition $A(t) = e^{i\lambda_0 t}a(t)$, we get from (4.9)

$$\begin{aligned} \phi_1(t) &= -\frac{i\varepsilon}{2}\int_0^t e^{-iH_0(t-s)}e^{-i\lambda_0 s}A(s)\mathbf{P_c}\sum_{j\in\mathbb{Z}} e^{-i\mu_j s}\ \beta_j\psi_0\ ds \\ &= -\frac{i\varepsilon}{2}\sum_{j\in\mathbb{Z}}\int_0^t e^{-iH_0(t-s)}e^{-i(\lambda_0+\mu_j)s}A(s)\mathbf{P_c}\beta_j\psi_0\ ds. \end{aligned} \tag{4.14}$$

We would like to integrate by parts each of the integrals in the above sum. We cannot proceed directly since the resolvents of H_0 in $\lambda_0 + \mu_j,\ j \in \mathbb{Z}$, would appear and hypothesis **(H5)** implies that some of the $\lambda_0 + \mu_j,\ j \in \mathbb{Z}$, are in the spectrum of H_0. Instead we regularize ϕ_1 by defining

$$\phi_1^\eta(t) = -\frac{i}{2}\varepsilon\sum_{j\in\mathbb{Z}}\int_0^t e^{-iH_0(t-s)}e^{-i(\lambda_0+\mu_j+i\eta)s}A(s)\mathbf{P_c}\beta_j\psi_0\ ds \tag{4.15}$$

for η positive and arbitrary and $t > 0$. Note that $\phi_1(t) = \lim_{\eta\searrow 0}\phi_1^\eta(t)$ uniformly with respect to t on compact intervals.

Now, integration by parts for each integral in expression (4.15) and letting η tend to zero from above gives the following expansion of $(\psi_0, W(t)\phi_1(t))$:

$$\begin{aligned} (\psi_0, W(t)\phi_1(t)) &= \left(W(t)\psi_0,\ -\frac{\varepsilon}{2}\ e^{-i\lambda_0 t}\sum_{j\in\mathbb{Z}} e^{-i\mu_j t}A(t)(H_0-\lambda_0-\mu_j-i0)^{-1}\mathbf{P_c}\beta_j\psi_0\right) \\ &+\left(W(t)\psi_0,\ \frac{\varepsilon}{2}A(0)\sum_{j\in\mathbb{Z}} e^{-iH_0 t}(H_0-\lambda_0-\mu_j-i0)^{-1}\mathbf{P_c}\beta_j\psi_0\right) \\ &+\left(W(t)\psi_0,\ \frac{\varepsilon}{2}\sum_{j\in\mathbb{Z}}\int_0^t e^{-iH_0(t-s)}(H_0-\lambda_0-\mu_j-i0)^{-1}\mathbf{P_c}e^{-i(\lambda_0+\mu_j)s}\partial_s A(s)\beta_j\psi_0 ds\right). \end{aligned} \tag{4.16}$$

The definition of the singular operators in the above computation is given in section 8. The choice of regularization, $+i\eta$, in (4.15) ensures that the latter two terms in the expansion of ϕ_1, (4.16), decay dispersively as $t \to +\infty$; see hypothesis **(H3)** and section 6. For $t < 0$, we replace $+i\eta$ with $-i\eta$ in (4.15).

To further expand the first series in (4.16) we use the identities (8.5). The proof of Proposition 4.1 is now completed by substitution of (8.5) in the expansion (4.16) for ϕ_1 and of the result into the second term of the sum in (4.10). □

In the next sections we estimate the remainder terms in (4.9) and (4.11).

5. Estimates on the bound state amplitude. Our strategy is as follows. Equations (4.9) and (4.11) comprise a dynamical system governing $\phi_d(t)$ and $a(t) = A(t)e^{-i\lambda_0 t}$, the solution of which is equivalent to the original equation (1.1). In this and in the following section we derive a coupled system of estimates for $A(t)$ and $\phi_d(t)$. This section is focused on obtaining estimates for the bound state amplitude $A(t)$ in terms of $\phi_d(t)$, while the following section is focused on obtaining dispersive estimates for $\phi_d(t)$ in terms of $A(t)$. We treat only the case $t > 0$ since the modifications for the case $t < 0$ are obvious. The coupled system of estimates shows that $A(t)$ decays in time, provided $\phi_d(t)$ is dispersively decaying and vice-versa. We exploit the assumed smallness of the perturbation εW to "close" the resulting inequalities and prove the decay of both $A(t)$ and $\phi_d(t)$.

The main difference from the strategy employed in [22] for the estimation of the bound state amplitude is related to the presence of *infinitely many* frequencies in the perturbation $W(t)$. In particular, one can have an *accumulation of resonances* in the continuous spectrum of H_0. We have two strategies for obtaining estimates for $A(t)$ which correspond to the use of hypotheses **(H1)**–**(H5)** (Theorems 2.1 and 2.2) or hypotheses **(H1)**–**(H6)** (Theorem 2.3). These strategies revolve around estimation of $\Re \int_0^t \rho(s)\, ds$, where ρ is given by (4.12). Hypothesis **(H6)**, which controls certain "small divisors" which arise from the clustering of frequencies, ensures that

$$\Re \int_0^t \rho(s)\ ds \ \leq\ C\ \varepsilon^2 |||W|||^2. \tag{5.1}$$

This, in turn, implies that the contribution of $\rho(t)$ in the size of $A(t)$ is of order $\varepsilon^2|||W|||^2$. Without hypothesis **(H6)** we carefully decompose $\rho(t)$ as

$$\rho(t) \ =\ \varepsilon^2\sigma(t) \ +\ \eta(t),$$

where $\sigma(t)$ is a real almost periodic function with mean, $M(\sigma)$, zero and $\Re \int_0^t \eta(s)\, ds \leq C\varepsilon^2||W|||^2$. As in the previous case, the contribution of the $\eta(t)$ in the size of $A(t)$ is of order $\varepsilon^2|||W|||^2$. On the other hand, $\sigma(t)$ competes with the damping term $\varepsilon^2\Gamma$ in (4.11), but being oscillatory (i.e., of mean zero) and of the same size as the damping it allows the latter to eventually dominate.

As the above discussion suggests it is simplest to start by assuming **(H6)** to get sharper estimates on $A(t)$ (Theorem 2.3) and then to relax this assumption (Theorem 2.2). We begin with a simple lemma which we shall use in a number of places in this and in the next section.

Lemma 5.1. *Let* $\alpha > 1$.

$$\int_0^t \langle t-s\rangle^{-\alpha}\ \langle s\rangle^{-\beta}\ ds \ \leq C_{\alpha,\beta}\ \langle t\rangle^{-\min(\alpha,\beta)}. \tag{5.2}$$

Proof. The bound is obtained by viewing the integral as decomposed into a part over $[0, t/2]$ and the part over $[t/2, t]$. We estimate the integral over $[0, t/2]$ by bounding $\langle t-s\rangle^{-\alpha}$ by its value at $t/2$ and explicitly computing the remaining integral. The integral over $[t/2, t]$ is computed by bounding $\langle s\rangle^{-\beta}$ by its value at $t/2$ and again computing explicitly the remaining integral. Putting the two estimates together yields the lemma.

We now turn to the estimate for $A(t)$ in terms of the dispersive norm of $\phi_d(t)$ and local decay estimates for $e^{-iH_0t}\mathbf{P_c}(H_0)$.

5.1. Estimates for $A(t)$ under the hypotheses of Theorem 2.3.

PROPOSITION 5.1. *Suppose* **(H1)**–**(H6)** *hold. Then $A(t)$, the solution of* (4.11), *can be expanded as*

$$A(t) \;=\; e^{\int_0^t \rho(s)ds}\left(e^{-\varepsilon^2\Gamma t}A(0) + R_A(t)\right), \tag{5.3}$$

$$R_A(t) \;=\; \int_0^t e^{-\varepsilon^2\Gamma(t-\tau)}\ \tilde{E}(\tau)\ d\tau, \tag{5.4}$$

where $\tilde{E}(t)$ is given in (4.13) *and* (5.9). *For any $\alpha > 1$, there exists a $\delta > 0$ such that $R_A(t)$ satisfies the estimates for $T > 2(\varepsilon^2\Gamma)^{-\alpha}$,*

$$\begin{aligned}\sup_{2(\varepsilon^2\Gamma)^{-\alpha}\le t\le T} \langle t\rangle^{r_1}\ |R_A(t)| \le C_1 e^{-(\varepsilon^2\Gamma)^{-\delta}} &\sup_{0\le\tau\le(\varepsilon^2\Gamma)^{-\alpha}} |E(\tau)| \\ &+ C\varepsilon^2\Gamma^{-1} \sup_{(\varepsilon^2\Gamma)^{-\alpha}\le\tau\le T} \left(\langle\tau\rangle^{r_1}|E(\tau)|\right),\end{aligned} \tag{5.5}$$

$$\sup_{0\le t\le 2(\varepsilon^2\Gamma)^{-\alpha}} \langle t\rangle^{r_1}\ |R_A(t)| \le\ D\ (\varepsilon^2\Gamma)^{-\alpha(r_1+1)} \sup_{0\le\tau\le 2(\varepsilon^2\Gamma)^{-\alpha}} |E(\tau)|. \tag{5.6}$$

Proof. To prove (5.5) we begin with (4.11). Let

$$\tilde{A}(t) \equiv e^{-\int_0^t \rho(s)ds}A(t). \tag{5.7}$$

Then, $\tilde{A}$ satisfies the equation

$$\partial_t\tilde{A} = -\varepsilon^2\Gamma\tilde{A} + \tilde{E}(t), \tag{5.8}$$

$$\tilde{E}(t) \equiv e^{-\int_0^t \rho(s)ds}E(t). \tag{5.9}$$

Solving (5.8) we get

$$\tilde{A}(t) \;=\; e^{-\varepsilon^2\Gamma t}\tilde{A}(0) \;+\; \int_0^t e^{-\varepsilon^2\Gamma(t-s)}\tilde{E}(s)\ ds \tag{5.10}$$

$$\equiv\; e^{-\varepsilon^2\Gamma t}\tilde{A}(0) \;+\; R_A(t). \tag{5.11}$$

Below, in Proposition 5.2 we show that the real part of the integral of $\rho(t)$ is uniformly bounded and of order $\mathcal{O}(\varepsilon^2|||W|||^2)$ for $t \ge 0$. Therefore, for some $C > 0$, we have by (5.7) and (5.9)

$$C^{-1}|\tilde{A}(t)| \le |A(t)| \le C|\tilde{A}(t)|, \tag{5.12}$$

$$C^{-1}|\tilde{E}(t)| \le |E(t)| \le C|\tilde{E}(t)|. \tag{5.13}$$

Consequently, it is sufficient to estimate $\tilde{A}(t)$, in terms of $\tilde{E}(t)$.

Remark 5.1. Estimates of $R_a(t)$, which appear in the statement of Theorem 2.3, are related to those for $R_A(t)$ via

$$R_a(t) \;=\; e^{-i\lambda_0 t+\int_0^t \rho(s)\ ds}\ R_A(t) - \left(1 - e^{\Re\int_0^t \rho(s)ds}\right)e^{-\varepsilon^2\Gamma t}e^{iw(t)}a(0). \tag{5.14}$$

Hence, by Proposition 5.2,

$$|R_a(t)| \ \leq \ C \ |R_A(t)| + \mathcal{O}(\varepsilon^2 |||W|||^2). \tag{5.15}$$

From (5.10) we have for any $M > 0$

$$\begin{aligned} |\tilde{A}(t)| \ &\leq |A(0)| e^{-\varepsilon^2 \Gamma t} + \int_0^M e^{-\varepsilon^2 \Gamma (t-s)} |\tilde{E}(s)| ds + \int_M^t e^{-\varepsilon^2 \Gamma (t-s)} |\tilde{E}(s)| \ ds \\ &= \ |A(0)| e^{-\varepsilon^2 \Gamma t} \ + \ I_1(t) \ + \ I_2(t). \end{aligned} \tag{5.16}$$

Set

$$M = (\varepsilon^2 \Gamma)^{-\alpha}, \quad \alpha > 1.$$

We now estimate the terms $I_1(t)$ and $I_2(t)$ in (5.16) for $2(\varepsilon^2\Gamma)^{-\alpha} \leq t \leq T$.

$$\begin{aligned} \langle t \rangle^{r_1} \ I_1(t) \ &= \ \langle t \rangle^{r_1} \ \int_0^M e^{-\varepsilon^2 \Gamma (t-s)} |\tilde{E}(s)| ds \\ &\leq \ \langle t \rangle^{r_1} e^{-\frac{1}{2}\varepsilon^2 \Gamma t} \cdot \int_0^M e^{-\varepsilon^2 \Gamma (\frac{1}{2} t - s)} \ ds \ \cdot \sup_{0 \leq \tau \leq (\varepsilon^2 \Gamma)^{-\alpha}} |\tilde{E}(\tau)| \\ &\leq \ \sup_{2(\varepsilon^2\Gamma)^{-\alpha} \leq t \leq T} \left(\langle t \rangle^{r_1} e^{-\frac{1}{2}\varepsilon^2 \Gamma t} \right) \cdot C (\varepsilon^2 \Gamma)^{-1} \cdot \sup_{0 \leq \tau \leq (\varepsilon^2 \Gamma)^{-\alpha}} |\tilde{E}(\tau)| \\ &\leq \ C e^{-(\varepsilon^2 \Gamma)^{-\delta}} \sup_{0 \leq \tau \leq (\varepsilon^2 \Gamma)^{-\alpha}} |\tilde{E}(\tau)| \end{aligned} \tag{5.17}$$

for some $\delta > 0$. Therefore,

$$\sup_{2(\varepsilon^2\Gamma)^{-\alpha} \leq t \leq T} \left(\langle t \rangle^{r_1} \ I_1(t) \right) \ \leq \ C e^{-(\varepsilon^2 \Gamma)^{-\delta}} \sup_{0 \leq \tau \leq (\varepsilon^2 \Gamma)^{-\alpha}} |\tilde{E}(\tau)|. \tag{5.18}$$

We estimate $I_2(t)$ on the interval $2(\varepsilon^2\Gamma)^{-\alpha} \leq t \leq T$ as follows:

$$\langle t \rangle^{r_1} \ I_2(t) \ \leq \ \langle t \rangle^{r_1} \ \int_{(\varepsilon^2\Gamma)^{-\alpha}}^t e^{-\varepsilon^2 \Gamma (t-s)} \langle s \rangle^{-r_1} \ ds \ \sup_{(\varepsilon^2\Gamma)^{-\alpha} \leq \tau \leq T} \left(\langle \tau \rangle^{r_1} \tilde{E}(\tau) \right). \tag{5.19}$$

The integral is now bounded above using the estimate

$$\langle t \rangle^{r_1} \ \int_{(\varepsilon^2\Gamma)^{-\alpha}}^t \ e^{-\varepsilon^2 \Gamma (t-s)} \ \langle s \rangle^{-r_1} \ ds \ \leq \ C (\varepsilon^2 \Gamma)^{-1}, \ t \geq 2(\varepsilon^2\Gamma)^{-\alpha}. \tag{5.20}$$

This gives

$$\sup_{2(\varepsilon^2\Gamma)^{-\alpha} \leq t \leq T} \langle t \rangle^{r_1} \ I_2(t) \ \leq \ C (\varepsilon^2 \Gamma)^{-1} \ \sup_{(\varepsilon^2\Gamma)^{-\alpha} \leq \tau \leq T} \left(\langle \tau \rangle^{r_1} \tilde{E}(\tau) \right). \tag{5.21}$$

Assembling the estimates (5.18) and (5.21) yields estimate (5.5) of Proposition 5.1 provided that (5.12) and (5.13) hold. Estimate (5.6) is a simple consequence of the definition of $R_A(t)$.

Thus it remains to prove (5.12) and (5.13). By (5.7) and (5.9) it is necessary and sufficient to verify the following proposition.

Proposition 5.2. *Assume hypotheses* **(H1)**–**(H6)**. *If ρ is given by* (4.12), *then*

$$\Re \int_0^t \rho(s)\ ds \le C\varepsilon^2 |||W|||^2,\ t \ge 0, \tag{5.22}$$

for some constant C depending on $\mathcal{C}$, r_1, and ξ; see **(H6)**.

Proof of Proposition 5.2. Using the estimates (8.7) and (8.9) we can infer that $\rho(t)$, given by (4.12) is a series which converges uniformly on any compact subset of $\mathbb{R}$. For each fixed t, it can therefore be integrated term-by-term to give

$$\begin{aligned}\Re \int_0^t \rho(s) ds &= \frac{\varepsilon^2}{4} \Re\ i \sum_{j,k\in\mathbb{Z}, j\neq k} \int_0^t e^{i(\mu_k-\mu_j)s} \left(\beta_k\psi_0, (H_0-\lambda_0-\mu_j-i0)^{-1}\mathbf{P_c}\beta_j\psi_0\right) ds \\ &= \frac{\varepsilon^2}{4} \sum_{j,k\in\mathbb{Z}, j\neq k} \Re \frac{e^{i(\mu_k-\mu_j)t}-1}{\mu_k-\mu_j} \left(\beta_k\psi_0, (H_0-\lambda_0-\mu_j-i0)^{-1}\mathbf{P_c}\beta_j\psi_0\right). \end{aligned} \tag{5.23}$$

Define

$$\tilde{\rho}_{j,k} \equiv \frac{e^{i(\mu_k-\mu_j)t}-1}{\mu_k-\mu_j} \left(\beta_k\psi_0, (H_0-\lambda_0-\mu_j-i0)^{-1}\mathbf{P_c}\beta_j\psi_0\right).$$

Then (5.23) can be expressed as

$$\Re \int_0^t \rho(s) ds = \frac{\varepsilon^2}{4} \sum_{j,k\in\mathbb{Z}, j\neq k} \Re\tilde{\rho}_{j,k} = \frac{\varepsilon^2}{8} \sum_{j,k\in\mathbb{Z}, j\neq k} \Re(\tilde{\rho}_{j,k}+\tilde{\rho}_{k,j}). \tag{5.24}$$

Now, since

$$\begin{aligned}\tilde{\rho}_{k,j} &= -\frac{e^{-i(\mu_k-\mu_j)t}-1}{\mu_k-\mu_j} \left(\beta_j\psi_0, (H_0-\lambda_0-\mu_k-i0)^{-1}\mathbf{P_c}\beta_k\psi_0\right) \\ &= -\left[\frac{e^{i(\mu_k-\mu_j)t}-1(\ldots)}{\mu_k-\mu_j}\right]^* \left(\beta_k\psi_0, (H_0-\lambda_0-\mu_k+i0)^{-1}\mathbf{P_c}\beta_j\psi_0,\right)\end{aligned}$$

we have

$$\begin{aligned}&\Re(\tilde{\rho}_{j,k}+\tilde{\rho}_{k,j}) \\ &= \Re \frac{e^{i(\mu_k-\mu_j)t}-1}{\mu_k-\mu_j} \left(\beta_k\psi_0, (H_0-\lambda_0-\mu_j-i0)^{-1} - (H_0-\lambda_0-\mu_k+i0)^{-1}\mathbf{P_c}\beta_j\psi_0\right).\end{aligned} \tag{5.25}$$

Moreover, by (8.5) we can infer

$$\Re(\tilde{\rho}_{j,k}+\tilde{\rho}_{k,j}) = \Re\left(e^{i(\mu_k-\mu_j)t}-1\right)\rho_{j,k} + 2\Im\left(e^{-i(\mu_k-\mu_j)t}-1\right)\delta_{j,k},$$

where, for $j \neq k \in \mathbb{Z}$,

$$\rho_{j,k} \equiv \frac{1}{\mu_k-\mu_j} \left(\beta_k\psi_0, (H_0-\lambda_0-\mu_j-i0)^{-1} - (H_0-\lambda_0-\mu_k-i0)^{-1}\mathbf{P_c}\beta_j\psi_0\right), \tag{5.26}$$

and for $j \neq k,\ j \in \mathbb{Z},\ k \in I_{res}$,

$$\delta_{j,k} \equiv \frac{\pi}{\mu_k-\mu_j} \left(\beta_j\psi_0, \delta(H_0-\lambda_0-\mu_k)\beta_k\psi_0\right). \tag{5.27}$$

Thus, by (5.24) and (5.25)

$$\Re\int_0^t \rho(s)ds = \frac{\varepsilon^2}{8}\sum_{j,k\in\mathbb{Z},j\neq k}\Re\big(e^{i(\mu_k-\mu_j)t}-1\big)\rho_{j,k}+\frac{\varepsilon^2}{4}\sum_{k\in I_{res},k\neq j\in\mathbb{Z}}\Im\big(e^{-i(\mu_k-\mu_j)t}-1\big)\delta_{j,k}. \tag{5.28}$$

We now derive a uniform bound for $\Re\ \int_0^t \rho(s)\ ds$.

Estimating the modulus of the above sum, we have for any t

$$\left|\ \Re\ \int_0^t \rho(s)\ ds\ \right|\ \leq\ \frac{\varepsilon^2}{4}\sum_{j,k\in\mathbb{Z},j\neq k}|\rho_{j,k}|+\frac{\varepsilon^2}{2}\sum_{k\in I_{res},k\neq j\in\mathbb{Z}}|\delta_{j,k}|\,. \tag{5.29}$$

By **(H6)**,

$$\sum_{k\in I_{res},k\neq j\in\mathbb{Z}}|\delta_{j,k}|\ \leq\ \pi\xi\ |||W|||^2. \tag{5.30}$$

We now bound the first term in (5.29). This requires an estimate of

$$|\rho_{j,k}| = \left|\frac{1}{\mu_k-\mu_j}\left(\beta_k\psi_0,(H_0-\lambda_0-\mu_j-i0)^{-1}-(H_0-\lambda_0-\mu_k-i0)^{-1}\mathbf{P_c}\beta_j\psi_0\right)\right|$$

for $j\neq k\in\mathbb{Z}$. We rely on the hypothesis **(H3b)** (singular local decay estimate (2.4)), which implies smoothness of the resolvent of H_0 near accumulation points in $\sigma_{\rm cont}(H_0)$ of the set $\{\lambda_0+\mu_j\}_{j\in\mathbb{Z}}$.

In order to treat both $\lambda_0+\mu_j\in\sigma_{\rm cont}(H_0)$ and $\lambda_0+\mu_j\notin\sigma_{\rm cont}(H_0)$ case simultaneously we regularize $\rho_{j,k}$:

$$\rho^\eta_{j,k}\equiv\frac{1}{\mu_k-\mu_j}\left(\beta_k\psi_0,(H_0-\lambda_0-\mu_j-i\eta)^{-1}-(H_0-\lambda_0-\mu_k-i\eta)^{-1}\mathbf{P_c}\beta_j\psi_0\right). \tag{5.31}$$

Clearly $\rho_{j,k}=\lim_{\eta\searrow 0}\rho^\eta_{j,k}$.

Now by the standard resolvent formula we have

$$\rho^\eta_{j,k}=\left(\beta_k\psi_0,(H_0-\lambda_0-\mu_k-i\eta)^{-1}(H_0-\lambda_0-\mu_j-i\eta)^{-1}\mathbf{P_c}\beta_j\psi_0\right).$$

Thus, using the singular local decay estimate **(H3b)**, we get

$$\begin{aligned}
|\rho_{j,k}| &= \left|\lim_{\eta\searrow 0}\int_0^\infty\left(\beta_k\psi_0,e^{-i(H_0-\lambda_0-\mu_k-i\eta)s}(H_0-\lambda_0-\mu_j-i\eta)^{-1}\mathbf{P_c}\beta_j\psi_0\right)ds\right|\\
&\leq\lim_{\eta\searrow 0}\int_0^\infty e^{-\eta s}\left|\left(w_+\beta_k\psi_0,w_-e^{-iH_0s}(H_0-\lambda_0-\mu_j-i\eta)^{-1}\mathbf{P_c}w_-w_+\beta_j\psi_0\right)\right|ds\\
&\leq\|w_+\beta_k\|\|w_+\beta_j\|\int_0^\infty\|w_-e^{-iH_0s}(H_0-\lambda_0-\mu_j-i0)^{-1}\mathbf{P_c}w_-\|ds\\
&\leq\mathcal{C}\|w_+\beta_k\|\|w_+\beta_j\|\int_0^\infty\langle s\rangle^{-r_1}ds\\
&\leq C\|w_+\beta_k\|\|w_+\beta_j\|
\end{aligned} \tag{5.32}$$

for some constant C depending on $\mathcal{C}$ and r_1. Summing on $j,k\in\mathbb{Z},\ j\neq k$, yields

$$\sum_{j,k\in\mathbb{Z},\ j\neq k}|\rho_{j,k}|\ \leq\ C\ |||W|||^2 \tag{5.33}$$

for some $C > 0$; see (2.7). Use of the bounds (5.30) and (5.33) in (5.29) gives

$$\left| \Re \int_0^t \rho(s)\ ds \right| \ \leq\ C\ \varepsilon^2 |||W|||^2$$

for some constant C depending on $\mathcal{C},\ r_1$, and ξ.

This completes the proof of Proposition 5.2 and therewith Proposition 5.1. □

5.2. Estimates for $\boldsymbol{A(t)}$ under the hypotheses of Theorem 2.1. In this subsection we work under the hypotheses of Theorem 2.1. In particular, we drop hypothesis **(H6)**. *We shall reuse the notation $\tilde{A}$ and $\tilde{E}$ for functions which are different from but related to those defined in section* 5.1.

PROPOSITION 5.3. *Suppose* **(H1)–(H5)** *hold. Then $A(t)$, the solution of* (4.11), *can be expanded as*

$$A(t) \ =\ e^{\int_0^t \eta(s)ds}\Big(e^{-\varepsilon^2(\Gamma t - \int_0^t \sigma(s)ds)} A(0) + R_A(t)\Big), \tag{5.34}$$

$$R_A(t) \ =\ \int_0^t \ e^{-\varepsilon^2\Gamma(t-\tau)+\varepsilon^2\int_\tau^t \sigma(s)ds}\ \tilde{E}(\tau)\ d\tau, \tag{5.35}$$

where

$$\sigma(t) \equiv -\frac{\pi}{4}\Re \sum_{j\in I_{res}, j\neq k\in\mathbb{Z}} e^{i(\mu_k-\mu_j)t}\,(\ \beta_k\psi_0, \delta(H_0-\lambda_0-\mu_j)\beta_j\psi_0) \tag{5.36}$$

is a real almost periodic function with mean $M(\sigma) = 0$, η in (5.46) *is a function whose real part has a bounded time integral of order $\mathcal{O}(\varepsilon^2|||W|||^2)$ and $\tilde{E}(t)$ is given in* (5.42)*; see also* (4.13). *For any $\alpha > 1$, there exists $\delta > 0$ such that $R_A(t)$ satisfies the estimates*

$$\begin{aligned} \sup_{2(\varepsilon^2\Gamma/2)^{-\alpha}\leq t\leq T} \langle t\rangle^{r_1}|R_A(t)| \leq C_1 e^{-(\varepsilon^2\Gamma/2)^{-\delta}} & \sup_{0\leq\tau\leq(\varepsilon^2\Gamma/2)^{-\alpha}} |E(\tau)| \\ +C(\varepsilon^2\Gamma)^{-1} & \sup_{(\varepsilon^2\Gamma/2)^{-\alpha}\leq\tau\leq T} \left(\langle\tau\rangle^{r_1}|E(\tau)|\right), \end{aligned} \tag{5.37}$$

$$\sup_{0\leq t\leq 2(\varepsilon^2\Gamma/2)^{-\alpha}} \langle t\rangle^{r_1}\ |R_A(t)| \leq\ D\ (\varepsilon^2\Gamma/2)^{-\alpha(r_1+1)} \sup_{0\leq\tau\leq 2(\varepsilon^2\Gamma/2)^{-\alpha}} |E(\tau)|. \tag{5.38}$$

Proof. As in the previous subsection we begin with the equation for $A(t)$:

$$\partial_t A(t) \ =\ \big(\rho(t)\ -\ \varepsilon^2\Gamma\ \big)\, A(t) + E(t), \tag{5.39}$$

where $\rho(t)$ and $E(t)$ are given by (4.12)–(4.13). In the previous section we transformed away the term $\rho(t)A(t)$ using the "integration factor" $\exp(\int_0^t \rho(s)\ ds)$. Under the current hypotheses, this can't be done because without **(H6)** $\Re\int_0^t \rho(s)\ ds$ may be unbounded as $t \to \infty$, which could cause the estimates (5.12)–(5.13) to break down. Instead, we proceed by a more refined analysis of $\rho(t)$, which we now outline.

We express $\rho(t)$ as $\rho(t) = \varepsilon^2\sigma(t) + \eta(t)$, where $\eta(t)$ has a time integral whose real part can be bounded by the estimates of section 5.1 and a part, $\varepsilon^2\sigma(t)$, which is almost periodic and of mean zero. Using this decomposition of $\rho(t)$ we write (5.39) as

$$\partial_t A(t) \ =\ \big[-\varepsilon^2\Gamma\ +\ \varepsilon^2\sigma(t)\ + \eta(t)\big]\, A(t)\ +\ E(t).$$

Next introduce the change of variables

$$\tilde{A}(t) \ \equiv \ e^{-\int_0^t \eta(s) \ ds} \ A(t) \tag{5.40}$$

and obtain a reduction to

$$\partial_t \tilde{A} = \left[\ -\varepsilon^2 \Gamma \ + \ \varepsilon^2 \sigma(t) \ \right] \tilde{A} \ + \ \tilde{E}(t), \tag{5.41}$$

$$\tilde{E}(t) \equiv e^{-\int_0^t \eta(s) ds} E(t). \tag{5.42}$$

With this strategy in mind we now proceed to derive the decomposition of $\rho(t)$. We are mostly interested in its real part, so we start with it.

$$\begin{aligned}
\Re\rho(t) &= \Re \frac{i\varepsilon^2}{4} \sum_{j,k\in\mathbb{Z},j\neq k} e^{i(\mu_k-\mu_j)t} \left(\beta_k\psi_0, (H_0-\lambda_0-\mu_j-i0)^{-1}\mathbf{P_c}\beta_j\psi_0\right) \\
&= -\frac{\varepsilon^2}{4} \Im \sum_{j,k\in\mathbb{Z},j\neq k} e^{i(\mu_k-\mu_j)t} \left(\beta_k\psi_0, (H_0-\lambda_0-\mu_j-i0)^{-1}\mathbf{P_c}\beta_j\psi_0\right) \\
&\equiv -\frac{\varepsilon^2}{4} \sum_{j,k\in\mathbb{Z},j\neq k} \Im\eta_{j,k} \\
&= \frac{\varepsilon^2}{8} \sum_{j,k\in\mathbb{Z},j\neq k} \Im\left(\eta_{j,k}+\eta_{k,j}\right).
\end{aligned} \tag{5.43}$$

In a manner similar to the derivation of (5.25) from (5.24) we find

$$\Im\eta_{k,j} = \Im e^{i(\mu_k-\mu_j)t} \left(\beta_k\psi_0, (H_0-\lambda_0-\mu_k+i0)^{-1}\mathbf{P_c}\beta_j\psi_0\right). \tag{5.44}$$

Using (8.5) in (5.44) and then replacing it in (5.43) we get

$$\begin{aligned}
\Re\rho(t) \ &= \ \frac{\pi}{4}\varepsilon^2 \Re \sum_{k\in I_{res},k\neq j\in\mathbb{Z}} e^{i(\mu_j-\mu_k)t} \left(\beta_j\psi_0, \delta(H_0-\lambda_0-\mu_k)\beta_k\psi_0\right) \\
&- \frac{1}{8}\varepsilon^2 \Im \sum_{j,k\in\mathbb{Z},j\neq k} e^{i(\mu_k-\mu_j)t} \big(\beta_k\psi_0, \big[(H_0-\lambda_0-\mu_j-i0)^{-1} \\
&\qquad\qquad -(H_0-\lambda_0-\mu_k-i0)^{-1}\big]\mathbf{P_c}\beta_j\psi_0\big) \\
&= \ \Re \ \eta(t) + \varepsilon^2\sigma(t).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\rho(t) \ &= \ \Re\rho(t) \ + \ i\Im\rho(t) \\
&= \ \eta(t) \ + \ \varepsilon^2\sigma(t),
\end{aligned} \tag{5.45}$$

where

$$\begin{aligned}
\eta(t) &= i \ \Im \ \rho(t) - \frac{1}{8}\varepsilon^2 \Im \sum_{j,k\in\mathbb{Z},j\neq k} e^{i(\mu_k-\mu_j)t} \\
&\quad \left(\beta_k\psi_0, \left[(H_0-\lambda_0-\mu_j-i0)^{-1} - (H_0-\lambda_0-\mu_k-i0)^{-1}\mathbf{P_c}\right]\beta_j\psi_0\right), \qquad (5.46) \\
\sigma(t) &= -\frac{\pi}{4}\Re \sum_{j\in I_{res},j\neq k\in\mathbb{Z}} e^{i(\mu_k-\mu_j)t} \left(\ \beta_k\psi_0, \delta(H_0-\lambda_0-\mu_j)\beta_j\psi_0\right);
\end{aligned}$$

see also (5.36).

Note that $\Re \int_0^t \eta(s)ds$ is uniformly bounded in t. To see this, recall the definition of $\rho_{j,k}$ in Lemma 5.2 (see (5.26)):

$$\rho_{j,k} \equiv \frac{1}{\mu_k - \mu_j} \left(\beta_k\psi_0, \left[(H_0 - \lambda_0 - \mu_j - i0)^{-1} - (H_0 - \lambda_0 - \mu_k - i0)^{-1}\right] \mathbf{P_c}\beta_j\psi_0\right).$$

By (8.7), $\Re\ \eta(t)$ given by (5.46) converges uniformly on $t \in \mathbb{R}$. Therefore, for each $t \in \mathbb{R}$ we may integrate the series term-by-term to obtain

$$\Re \int_0^t \eta(s)ds = \frac{1}{8}\varepsilon^2 \sum_{j,k\ j\neq k} \Re(e^{i(\mu_k-\mu_j)t} - 1)\rho_{j,k}. \tag{5.47}$$

Moreover, the modulus of the right-hand side in (5.47) is less or equal than $\frac{1}{4}\varepsilon^2 \sum_{j,k\ j\neq k} |\rho_{j,k}|$, which by (5.32) is bounded by $C\varepsilon^2|||W|||^2$ for some constant C depending only on $\mathcal{C}$ and r_1. Note that we derived (5.32) by using only hypothesis **(H3b)** and not relying on **(H6)**.

Thus we have

$$\Re \int_0^t \eta(s)ds \leq C\varepsilon^2|||W|||^2. \tag{5.48}$$

To summarize, we have split $\rho(t)$ into

$$\rho(t) = \eta(t)\ +\ \varepsilon^2\sigma(t)$$

such that (5.48) is valid. If we now define $\tilde{A}$ as in (5.40), then by (4.11) $\tilde{A}$ satisfies (5.41). Solving (5.41) we get

$$\begin{aligned} \tilde{A}(t) &= e^{-\varepsilon^2\Gamma t+\varepsilon^2\int_0^t \sigma(s)ds}\tilde{A}(0)\ +\ \int_0^t e^{-\varepsilon^2\Gamma(t-\tau)+\varepsilon^2\int_\tau^t \sigma(s)ds}\tilde{E}(s)\ d\tau \\ &\equiv e^{-\varepsilon^2\Gamma t+\varepsilon^2\int_0^t \sigma(s)ds}\tilde{A}(0)\ +\ R_A(t). \end{aligned} \tag{5.49}$$

From (5.42) and (5.48) it is sufficient to estimate $R_A(t)$ in terms of $\tilde{E}(t)$.

Remark 5.2. The estimates of $R_a(t)$ which appear in the statement of Theorem 2.2 are related to those for $R_A(t)$ via

$$R_a(t)\ =\ e^{-i\lambda_0 t+\int_0^t \eta(s)ds}R_A(t) + \left(1 - e^{\varepsilon^2(\int_0^t \sigma(s)ds-\gamma t)+\Re\int_0^t \eta(s)ds}\right)e^{-\varepsilon^2(\Gamma-\gamma)t}e^{iw(t)}a(0). \tag{5.50}$$

Before we estimate $R_A(t)$, we review some properties of the function $\sigma(t)$.

The function $\sigma(t)$ is almost periodic since the sum of the moduli of its Fourier coefficients is finite. Namely, by (2.20), the terms in the series (5.36) defining $\sigma(t)$ are majorized by those of a convergent series (whose sum is $C\pi^{-1}|||W|||^2$). Therefore, the series in (5.36) is uniformly convergent. As the uniform limit of almost periodic functions, $\sigma(t)$ is then itself almost periodic, bounded by

$$\sup_{t\in\mathbb{R}} |\sigma(t)|\ \leq\ C|||W|||^2 \tag{5.51}$$

for some constant C; see also section 9. Moreover, $\sigma(t)$ has mean value zero since all the Fourier exponents are nonzero; see (5.36) and section 9. Therefore

$$\int_\tau^t \sigma(s)ds \leq \frac{\Gamma}{2}(t-\tau),\ \text{for}\ \ t-\tau \geq \mathcal{M} \tag{5.52}$$

provided $\mathcal{M}$ is taken sufficiently large. It can be shown (see section 9 or [2, p. 42]) that (5.52) holds provided

$$\mathcal{M} \geq \frac{4\ \sup_{t\in\mathbb{R}}\{|\sigma(t)|\}\ L(\Gamma/4)}{\Gamma/2}, \tag{5.53}$$

where $L(\Gamma/4)$ (see Definition 9.1) is such that in each interval of length $L(\Gamma/4)$ there is at least one $\Gamma/4$ almost period for σ.

Using (5.51) and then **(H5)**, we can choose

$$\mathcal{M} = 8CL(\Gamma/4)/\theta_0 \tag{5.54}$$

independently of ε and still satisfy (5.53).

We now return to the estimation of R_A. We split the integral in (5.35) into two integrals, one from 0 to $t-\mathcal{M}$ and the other from $t-\mathcal{M}$ to t. For the former we use (5.52) while for the latter we use (5.51). The result is

$$\begin{aligned} |R_A(t)| \ &\leq \int_0^{t-\mathcal{M}} e^{-\frac{1}{2}\varepsilon^2\Gamma(t-\tau)}|\tilde{E}(\tau)|d\tau \\ &\quad + \int_{t-\mathcal{M}}^{t} e^{\varepsilon^2(C|||W|||^2-\Gamma)(t-\tau)}|\tilde{E}(\tau)|d\tau. \end{aligned} \tag{5.55}$$

The first integral in (5.55) can be bounded exactly as the term $\int_0^t e^{-\varepsilon^2\Gamma(t-\tau)}|\tilde{E}(\tau)|d\tau$ in the proof of Proposition 5.1. The second integral in (5.55) is bounded in the following manner:

$$\begin{aligned} &\langle t\rangle^{r_1}\int_{t-\mathcal{M}}^{t} e^{\varepsilon^2(C|||W|||^2-\Gamma)(t-\tau)}|\tilde{E}(\tau)|d\tau \\ &\leq \frac{\langle t\rangle^{r_1}}{\langle t-\mathcal{M}\rangle^{r_1}}\int_{t-\mathcal{M}}^{t} e^{\varepsilon^2(C|||W|||^2-\Gamma)(t-\tau)}d\tau \sup_{t-\mathcal{M}\leq\tau\leq t}\left(\langle\tau\rangle^{r_1}|E(\tau)|\right) \\ &\leq D \sup_{(\varepsilon^2\Gamma/2)^{-\alpha}\leq\tau\leq t}\left(\langle\tau\rangle^{r_1}|E(\tau)|\right). \end{aligned} \tag{5.56}$$

Note that ε and consequently $\varepsilon^2\Gamma \sim \varepsilon^2|||W|||^2$ are small, so we can consider $\mathcal{M} \ll (\varepsilon^2\Gamma/2)^{-\alpha}$ and $D \sim \mathcal{M} \ll (\varepsilon^2\Gamma)^{-1}$. The result is (5.37). A simple bound, using the definition of $R_A(t)$, yields (5.38).

This completes the proof of Proposition 5.3.

6. Dispersive estimates and local decay. In this section we prove the local decay of ϕ_d and the decay in time of the remainder terms, $E(t)$, in bound state amplitude equation (4.11) of section 4. The arguments rely on hypotheses **(H1)**–**(H5)** and results of the previous section, so we will handle Theorem 2.1 first. However, due to the differences between Theorems 2.2 and 2.3 we separately finish their proofs in the final two subsections of this section. We will repeatedly use the following lemma.

LEMMA 6.1. *For any $\eta \in [0, r_1]$ and $j \in \mathbb{Z}$ we have*

$$\left\|\int_0^t w_- e^{-iH_0(t-s)}\mathbf{P_c} f(s)ds\right\| \ \leq\ C\langle t\rangle^{-\eta}\sup_{0\leq\tau\leq t}\left(\langle\tau\rangle^{\eta}\|w_+ f(\tau)\|\right) \tag{6.1}$$

and

$$\left\|\int_0^t w_- e^{-iH_0(t-s)}\mathbf{P_c}(H_0-\lambda_0-\mu_j-i0)^{-1}f(s)\ ds\right\| \ \leq\ C\langle t\rangle^{-\eta}\sup_{0\leq\tau\leq t}\left(\langle\tau\rangle^{\eta}\ \|w_+ f(\tau)\|\right). \tag{6.2}$$

Proof. The proof follows from the assumed local decay estimates on e^{-iH_0t}; see **(H3a)**. Namely, using that $r_1 > 1$,

$$\begin{aligned}\left\|\int_0^t w_- e^{-iH_0(t-s)}\mathbf{P_c} f(s)\ ds\right\| &\le \int_0^t \|w_- e^{-iH_0(t-s)}\ \mathbf{P_c} w_-\|_{\mathcal{L}(\mathcal{H})}\langle s\rangle^{-\eta}\ ds \\ &\quad \cdot \sup_{0\le\tau\le t}\left(\langle\tau\rangle^{\eta}\|w_+ f(\tau)\|\right) \\ &\le C\int_0^t \langle t-s\rangle^{-r_1}\langle s\rangle^{-\eta}\ ds\ \sup_{0\le\tau\le t}\left(\langle\tau\rangle^{\eta}\|w_+ f(\tau)\|\right) \\ &\le C\langle t\rangle^{-\eta}\sup_{0\le\tau\le t}\left(\langle\tau\rangle^{\eta}\,\|w_+ f(\tau)\|\right),\end{aligned}$$

which proves (6.1). The proof of (6.2) is identical and uses the singular local decay estimate of **(H3b)**. □

We now define the norms

$$[A]_\alpha(T) \ = \ \sup_{0\le\tau\le T}\langle\tau\rangle^{\alpha}|A(\tau)| \tag{6.3}$$

and

$$[\phi_d]_{LD,\alpha}(T) \ = \ \sup_{0\le\tau\le T}\langle\tau\rangle^{\alpha}\|w_-\phi_d(\tau)\|. \tag{6.4}$$

Then we have the following.

Proposition 6.1. *For any $T > 0$ and $\eta \in [0,\ r_1]$,*

$$[\phi_d]_{LD,\eta}(T) \le C\left(\ \|w_+\phi_d(0)\|\ +\ |\varepsilon|\ |||W|||\ [A]_\eta(T)\ \right). \tag{6.5}$$

Proof. From (4.9) we get, using the assumed local decay estimate for e^{-iH_0t} and (6.1),

$$\begin{aligned}\|w_-\phi_d(t)\| &\le \sum_{j=0}^{2}\ \|w_-\phi_j(t)\| \\ &\le C\langle t\rangle^{-\eta}\,\|w_+\phi_d(0)\| + \ C|\varepsilon|\langle t\rangle^{-\eta}\ [A]_\eta(t)\ \sup_{0\le s\le t}\|w_+W(s)\psi_0\| \\ &\quad + C\ |\varepsilon|\ |||W|||\ \langle t\rangle^{-\eta}\,[\phi_d]_{LD,\eta}(t).\end{aligned} \tag{6.6}$$

Since $\|w_+W(s)\psi_0\| \le |||W|||\ \|\psi_0\| \ = \ |||W|||$ and $|\varepsilon|\ |||W|||$ is assumed to be small, multiplying both sides of this last equation by $\langle t\rangle^{\eta}$ and taking supremum over $t \le T$ yields (6.5). □

We now estimate $E(t)$.

Proposition 6.2. *Let $T > 0$. For any $\eta \in [0, r_1]$*

$$[E]_\eta(T)\ \le\ C\left(\varepsilon^2|||W|||^2\ |A(0)|\ +\ |\varepsilon|\ |||W|||\ \|w_+\phi_d(0)\|\ +\ |\varepsilon|^3|||W|||^3\ [A]_\eta(T)\right). \tag{6.7}$$

Proof. $E(t)$ is defined in (4.13). From these equations it is seen that we need to bound the following terms:

$$R_1 \equiv \frac{1}{4}\varepsilon^2|A(0)|\sum_{j,k\in\mathbb{Z}}\left|\left(\beta_k\psi_0,\ e^{-iH_0t}\ (H_0-\lambda_0-\mu_j-i0)^{-1}\mathbf{P_c}\beta_j\psi_0\right)\right|,$$

$$R_2 \equiv \frac{1}{4}\varepsilon^2\sum_{j,k\in\mathbb{Z}}\left|\left(\beta_k\psi_0,\int_0^t\ e^{-iH_0(t-s)}(H_0-\lambda_0-\mu_j-i0)^{-1}\mathbf{P_c}e^{-i(\lambda_0+\mu_j)s}\partial_s A(s)\beta_j\psi_0 ds\right)\right|$$

and

$$\begin{aligned}|\varepsilon\left(\psi_0,\ W(t)\phi_0(t)\right)| &= \left|\varepsilon\left(W(t)\psi_0, e^{-iH_0 t}\phi_d(0)\right)\right|,\\ |\varepsilon\left(\psi_0,\ W(t)\phi_2\right)| &= \left|\varepsilon^2\left(W(t)\psi_0, \int_0^t e^{-iH_0(t-s)}\mathbf{P_c}W(s)\phi_d(s)ds\right)\right|.\end{aligned}$$

The estimates of the above terms repeatedly use Lemma 6.1. Let $\eta \in [0, r_1]$.

Estimation of R_1.

$$\begin{aligned}R_1 &= \frac{1}{4}\varepsilon^2|A(0)|\sum_{j,k\in\mathbb{Z}}\left|\left(w_+\beta_k\psi_0,\ w_-e^{-iH_0 t}(H_0-\lambda_0-\mu_j-i0)^{-1}\mathbf{P_c}w_-\ w_+\beta_j\psi_0\right)\right|\\ &\le C|A(0)|\ \varepsilon^2|||W|||^2\ \langle t\rangle^{-\eta} \qquad (6.8)\end{aligned}$$

by the local decay estimates (2.4).

Estimation of R_2. From (4.11) we have that

$$|\partial_s A(s)| \le C|\varepsilon|\ |||W|||\ |A(s)| + |E(s)| \qquad (6.9)$$

since $\Im\rho$ is linear in $|\varepsilon|\ |||W|||$ and $\Re\rho$, Γ are quadratic.

Applying Lemma 6.1 to R_2 we then get

$$\begin{aligned}R_2 &= \tfrac{1}{4}\varepsilon^2\textstyle\sum_{j,k\in\mathbb{Z}}\left|\left(w_+\beta_k\psi_0, \int_0^t w_-e^{-iH_0(t-s)}(H_0-\lambda_0-\mu_j-i0)^{-1}\mathbf{P_c}w_-\partial_s A(s)w_+\beta_j\psi_0 ds\right)\right|\\ &\le\ C\varepsilon^2|||W|||^2\ \langle t\rangle^{-\eta}\big(|\varepsilon|\ |||W|||\ [A]_\eta(t)\ +\ [E]_\eta(t)\big). \qquad (6.10)\end{aligned}$$

Estimation of $|\varepsilon\left(\psi_0,\ W(t)\phi_0(t)\right)|$. Since, by definition, $\phi_d(0) = \mathbf{P_c}\phi_d(0)$ we can apply local decay estimates for $e^{-iH_0 t}$ to get

$$|\varepsilon\left(W(t)\psi_0,\ \phi_0(t)\right)| \le\ C|\varepsilon|\ |||W|||\ \langle t\rangle^{-\eta}\ \|w_+\phi_d(0)\|. \qquad (6.11)$$

Estimation of $|\varepsilon\left(\psi_0,\ W(t)\phi_2\right)|$. Applying Lemma 6.1 as before we get, for $0 \le t \le T$,

$$|\varepsilon\left(\psi_0,\ W(t)\phi_2\right)| \le C\varepsilon^2|||W|||^2\ \langle t\rangle^{-\eta}\ [\phi_d]_{LD,\eta}(T). \qquad (6.12)$$

Using Proposition 6.1 to estimate $[\phi_d]_{LD,\eta}(t)$ in (6.12), we get

$$|\varepsilon\left(\psi_0,\ W(t)\phi_2\right)| \le C\varepsilon^2|||W|||^2\ \langle t\rangle^{-\eta}\left\{\|w_+\phi_d(0)\|\ +\ |\varepsilon|\ |||W|||\ [A]_\eta(t)\right\}. \qquad (6.13)$$

Finally, combining the above estimates, we can bound $[E]_\eta(T)$ for any $\eta \in [0, r_1]$ as follows:

$$\begin{aligned}[E]_\eta(T) \le C\big\{&\varepsilon^2\ |||W|||^2\ |A(0)| + |\varepsilon|\ |||W|||\ \|w_+\phi_d(0)\|\\ &+\ \varepsilon^2|||W|||^2\ [E]_\eta(T) + |\varepsilon|^3|||W|||^3\ [A]_\eta(T)\big\}. \qquad (6.14)\end{aligned}$$

Since $|\varepsilon|\ |||W|||$ is assumed to be small, Proposition 6.2 follows. □

We can now complete the proof of Theorem 2.1. To prove the assertions concerning the infinite time behavior, the key is to establish local decay of ϕ_d, in particular, the uniform boundedness of $[\phi_d]_{LD,r_1}(T)$. This will follow directly from Proposition 6.1 if we prove the uniform boundedness $[A]_{r_1}(T)$, or equivalently, $[\tilde{A}]_{r_1}(T)$.

PROPOSITION 6.3. *Under the hypothesis of Theorem* 2.1, *there exists an* $\varepsilon_0 > 0$ *such that for each real number* ε, $|\varepsilon| < \varepsilon_0$ *there is a constant* C_* *with the property that for any* $T > 0$

$$[A]_{r_1}(T) \le C_*.$$

Proof. We begin with the expansion of $A(t)$ given in Proposition 5.3. Multiplying (5.34) by $\langle t\rangle^{r_1}$, and taking the supremum over $0 \le t \le T$ we have

$$\begin{aligned}[A]_{r_1}(T) \ \le\ C\Bigg(& |A(0)|\ (\varepsilon^2\Gamma/2)^{-r_1} \ + \sup_{0\le\tau\le 2(\varepsilon^2\Gamma/2)^{-\alpha}} \langle\tau\rangle^{r_1}|R_A(\tau)| \\ & + \sup_{2(\varepsilon^2\Gamma/2)^{-\alpha}\le\tau\le T} \langle\tau\rangle^{r_1}|R_A(\tau)|\ \Bigg).\end{aligned} \tag{6.15}$$

The right-hand side of (6.15) is estimated using Proposition 5.3.

$$\begin{aligned}[A]_{r_1}(T) \le C|A(0)|\ (\varepsilon^2\Gamma/2)^{-r_1} \ + D\ (\varepsilon^2\Gamma/2)^{-\alpha(r_1-1)}\ [E]_0(2(\varepsilon^2\Gamma/2)^{-\alpha}) \\ +\ C_1\ e^{-(\varepsilon^2\Gamma/2)^{-\delta}}\ [E]_0(2(\varepsilon^2\Gamma/2)^{-\alpha})\ + C_2\ (\varepsilon^2\Gamma/2)^{-1}[E]_{r_1}(T).\end{aligned}$$

Next, we apply Proposition 6.2 which yields

$$\begin{aligned}[A]_{r_1}(T) \le C|A(0)|\ (\varepsilon^2\Gamma/2)^{-r_1}\ +\ D\ (\varepsilon^2\Gamma/2)^{-\alpha(r_1+1)}\ [E]_0(2(\varepsilon^2\Gamma/2)^{-\alpha}) \\ +\ C_1 e^{-(\varepsilon^2\Gamma/2)^{-\delta}}\ [E]_0(2(\varepsilon^2\Gamma/2)^{-\alpha}) \\ +\ C_2(\varepsilon^2\Gamma/2)^{-1}\big(\varepsilon^2|A(0)||||W|||^2\ +\ |\varepsilon|\ |||W|||\ \|w_+\phi_d(0)\| \\ +\ |\varepsilon|^3|||W|||^3\ [A]_{r_1}(T)\big).\end{aligned} \tag{6.16}$$

Note that by Proposition 6.2 and the simple bound

$$[A]_0(T)\ \le\ \|\phi_0\|,$$

$[E]_0(2(\varepsilon^2\Gamma/2)^{-\alpha})$ is bounded in terms of the initial data and $|\varepsilon|\ |||W|||$.

Choose ε_0 such that

$$1 - \frac{2C_2|||W|||^3}{\Gamma}\varepsilon_0 = 0,$$

where C_2 is the same as in (6).

Then, for $|\varepsilon| < \varepsilon_0$

$$[A]_{r_1}(T) \le C_*. \tag{6.17}$$

Here, C_* depends on $\|\phi_0\|$, $\|w_+\phi_0\|$, r_1, and ε.

This completes the proof of Proposition 6.3 and therewith the $t \to \infty$ asymptotics asserted in Theorems 2.1–2.3. □

It remains to finish the proofs of the Theorems 2.3 and 2.2. Due to some differences we consider them separately in the following two subsections.

6.1. Proof of Theorem 2.3. In order to obtain (2.23) we note that (4.7), (5.3), and (5.14) together with the definition of $\omega(t)$ in (2.16) already gives us

$$\begin{aligned} a(t) &= e^{-i\lambda_0 t+\int_0^t \rho(s)ds}\big(A(0)e^{-\varepsilon^2\Gamma t} + R_A(t)\big) \\ &= a(0)e^{-\varepsilon^2\Gamma t}e^{i\omega(t)} + R_a(t), \end{aligned}$$

which is in fact the second relation in (2.23). The third is a direct consequence of the second since $P(t) = |a(t)|^2$ while the fourth relation is exactly (4.9).

It remains to prove the intermediate time estimate (2.17). The ingredients are contained in (6.7) and its proof. First, by (5.15)

$$|R_a(t)| \ \le C\ |R_A(t)| \ +\ \mathcal{O}(\varepsilon^2|||W|||^2).$$

So, it suffices to prove an $\mathcal{O}(|\varepsilon|\ |||W|||)$ upper bound for R_A.

Using (5.4) and (5.13) we know that

$$|R_A(t)| \le C\int_0^t\ e^{-\varepsilon^2\Gamma(t-\tau)}\ \ |E(\tau)|\ \ d\tau. \tag{6.18}$$

Let T_0 denote an arbitrary fixed positive number. We estimate the equation (6.18) for $t \in [0, T_0(\varepsilon^2\Gamma)^{-1}]$. We bound the exponential in the integrand by one (explicit integration would give something of order $(\varepsilon^2\Gamma)^{-1}$) and bound $|E(\tau)|$ by estimating the expressions in the proof of Proposition 6.2. First, the estimates of Proposition 6.2 for R_1 and $|\varepsilon\,(\psi_0,\ W(t)\phi_0(t))|$ are useful as is. Integration of the bounds (6.8) and (6.11) gives

$$\int_0^t\ e^{-\varepsilon^2\Gamma(t-\tau)}R_1\ d\tau \le C\ \varepsilon^2|||W|||^2\ \|w_+\phi(0)\|,$$

$$\int_0^t\ e^{-\varepsilon^2\Gamma(t-\tau)}\ |\varepsilon\,(\psi_0,\ W(t)\phi_0(t))|\ \ d\tau \le C\ |\varepsilon|\ \ |||W|||\ \|w_+\phi(0)\|. \tag{6.19}$$

To estimate the contributions of R_2, first observe that by (6.9) and Proposition 6.2 with $\eta = 0$

$$|\partial_s A(s)|\ \ \le\ \ C\ |\varepsilon|\ \ |||W|||\ \|w_+\phi(0)\|. \tag{6.20}$$

Therefore, using local decay estimates we have

$$\begin{aligned} \int_0^t\ e^{-\varepsilon^2\Gamma(t-\tau)}R_2\ d\tau \le&\ \ C\ T_0(\varepsilon^2\Gamma)^{-1}\ |\varepsilon|^3|||W|||^3\ \|w_+\phi(0)\| \\ \le&\ \ D|\varepsilon|\ \ |||W|||\ \|w_+\phi(0)\|. \end{aligned}$$

Finally, we come to the contribution of $|\varepsilon\,(\psi_0,\ W(t)\phi_2)|$. We rewrite it as follows:

$$\begin{aligned} |\varepsilon\,(\psi_0,\ W(t)\phi_2)| =&\ \ \varepsilon^2\left|\int_0^t\ \big(W(s)e^{iH_0(t-s)}\mathbf{P_c}W(t)\psi_0, \phi_d(s)\big)ds\right| \\ =&\ \ \left|\int_0^t\ \varepsilon^2(w_+W(s)w_+\cdot w_-e^{iH_0(t-s)}\mathbf{P_c}w_-\cdot w_+W(t)\psi_0,\ w_-\phi_d(s))\ ds\right|. \end{aligned} \tag{6.21}$$

Recall that by (4.9) $\phi_d\ =\ \phi_0\ +\ \phi_1\ +\ \phi_2$, where $\phi_0(t) = e^{-iH_0 t}\phi_d(0)$. Using local decay estimates **(H3a)**, the contribution of the term $\phi_0(t)$ can be bounded by $C\ \varepsilon^2|||W|||^2\ \|w_+\phi_d(0)\|\ \langle\tau\rangle^{-r_1}$. Multiplication of this bound by $e^{-\varepsilon^2\Gamma(t-\tau)}$ and integration with respect to t gives the bound $C\ \varepsilon^2|||W|||^2\|w_+\phi_d(0)\|$. To assess the contributions from $\phi_1+\phi_2$, note that local decay estimates **(H3a)** imply

$$\|w_-(\phi_1+\phi_2)\|\ \le\ C\ |\varepsilon|\ |||W|||\ \|w_+\phi(0)\|. \tag{6.22}$$

Putting together the contributions from ϕ_0 and from $\phi_1+\phi_2$, we have

$$\int_0^t e^{-\varepsilon^2\Gamma(t-\tau)}\left|\varepsilon\left(\psi_0,\ W(t)\phi_2\right)\right|\,d\tau\ \le\ C\left(\ \varepsilon^2|||W|||^2\ \|w_+\phi_d(0)\|\ +\ (\varepsilon^2\Gamma)^{-1}\ |\varepsilon|^3|||W|||^3\right). \tag{6.23}$$

The above estimates and (5.15) imply (2.17). Now, (2.18) is a direct consequence of (2.17) and the relation $P(t)=|a(t)|^2$.

This concludes the proof of Theorem 2.3.

6.2. Proof of Theorem 2.2. As in the proof of Theorem 2.3 relations (4.7), (5.34), (5.50), and the definition of $\omega(t)$ in (2.16) gives

$$\begin{aligned} a(t) &= e^{-i\lambda_0 t+\int_0^t\eta(s)ds}\left(A(0)e^{-\varepsilon^2\left(\Gamma t-\int_0^t\sigma(s)ds\right)}+R_A(t)\right)\\ &= a(0)e^{-\varepsilon^2(\Gamma-\gamma)t}e^{i\omega(t)}+R_a(t), \end{aligned}$$

which is the second relation in (2.15). In what follows, the only difference from the previous argument is in estimating $R_a(t)$.

We start with the relation (5.50):

$$R_a(t)\ =\ e^{-i\lambda_0 t+\int_0^t\eta(s)ds}R_A(t)+\left(1-e^{\varepsilon^2(\int_0^t\sigma(s)ds-\gamma t)+\Re\int_0^t\eta(s)ds}\right)e^{-\varepsilon^2(\Gamma-\gamma)t}e^{iw(t)}a(0). \tag{6.24}$$

Since $\sigma(t)$ is an almost periodic function with zero mean, for any $\gamma>0$ there is an $\mathcal{M}_\gamma>0$ such that whenever $|t|\ge\mathcal{M}_\gamma$

$$\int_0^t\sigma(s)ds\le\gamma t.$$

On the other hand for $|t|<\mathcal{M}_\gamma$, using (5.51) we have

$$\int_0^t\sigma(s)ds\le C\mathcal{M}_\gamma|||W|||^2.$$

So, in both cases,

$$\int_0^t\sigma(s)ds-\gamma t\le C\mathcal{M}_\gamma|||W|||^2.$$

Substituting now in (6.24) and tacking into account that by (5.48),

$$\Re\int_0^t\eta(s)ds\ \le\ C\varepsilon^2|||W|||^2$$

uniformly in t, we get

$$|R_a(t)|\ \le\ C\,|R_A(t)|+\mathcal{O}(\varepsilon^2|||W|||^2). \tag{6.25}$$

It remains to prove an $\mathcal{O}(|\varepsilon|\ |||W|||)$ for $R_A(t)$. Looking now at (5.55) we see that we can bound the exponential by $\max\{1,e^{\varepsilon^2(C|||W|||^2-\Gamma)\mathcal{M}}\}$. Now, the same argument as in the end of the previous subsection will give us the required result.

This completes the proof of Theorem 2.2.

7. Generalizations. In the previous sections we considered perturbations of the form $\varepsilon W(t)$, with $W(t)$ independent of ε. In this section, we shall extend our theory to a more general class of potentials, W_ε, which are small for small ε but which may deform nontrivially as ε varies.

Consider a family of perturbations $\mathcal{W}$ and the general system

$$\begin{aligned} i\partial_t \phi(t) &= \left(H_0 + W(t)\right)\phi(t), \\ \phi|_{t=0} &= \phi(0), \end{aligned} \tag{7.1}$$

where $W \in \mathcal{W}$ (compare to (2.1). The results are as follows.

THEOREM 7.1. *Suppose that H_0 and any $W \in \mathcal{W}$ satisfy hypotheses* **(H1)–(H5)**. *In addition assume the following:*

(H7) *Equi-almost periodicity. There exists a positive constant L_{θ_0}, independent of $W \in \mathcal{W}$, such that in any interval of real numbers of length L_{θ_0}, the function $|||W|||^{-2}\ \sigma(t)$ ($|||W||| \neq 0$), where*

$$\sigma(t) \equiv -\frac{\pi}{4}\Re \sum_{j \in I_{res}, j\neq k \in \mathbb{Z}} e^{i(\mu_k-\mu_j)t}\left(\beta_k\psi_0, \delta(H_0-\lambda_0-\mu_j)\beta_j\psi_0\right) \tag{7.2}$$

has a $\theta_0/4$ almost period, θ_0 is given by **(H5)**. *More precisely, there exists $L_{\theta_0} > 0$ which does not depend on W such that in any interval of length L_{θ_0} there is a number $\tau = \tau(\theta_0/4)$ such that for all $t \in \mathbb{R}$*

$$\left|\ |||W|||^{-2}\sigma(t+\tau)\ -\ |||W|||^{-2}\sigma(t)\ \right|\ \leq\ \theta_0/4. \tag{7.3}$$

If $w_+\phi(0) \in \mathcal{H}$, then there exists an $\varepsilon_0 > 0$ (depending on $\mathcal{C},\ r_1,\ \theta_0,$ and L_{θ_0}) such that whenever $|||W||| < \varepsilon_0$, the solution of (7.1) *satisfies the local decay estimate (identical with the one in Theorem* 2.1)

$$\|w_-\ \phi(t)\| \leq C\langle t\rangle^{-r_1}\|w_+\ \phi_0\|, \quad t \in \mathbb{R}. \tag{7.4}$$

Sketch of the proof. Once we drop ε from all expressions (since it is not present in the actual setting), the arguments in the previous sections hold in this case except the analysis of $\sigma(t)$ in Proposition 5.3. Formulas (5.36) and (7.2) are the same, but now $\mu_j,\ \beta_j,\ j \in \mathbb{Z}$, are not fixed as they define W by **(H4)** and W sweeps a general class $\mathcal{W}$. This may prevent us from finding a fixed time interval, $\mathcal{M}$, independent of $W \in \mathcal{W}$, after which $\sigma(t)$ is within $\Gamma/2$ distance from its mean; see relations (5.52)–(5.54).

Nevertheless, **(H7)** is exactly what we need to overcome the difficulty. A straightforward calculation shows that any $\theta_0/4$ almost period of $|||W|||^{-2}\sigma(t)$ is a $\Gamma/4$ almost period for $\sigma(t)$. Consequently, $L(\Gamma/4)$ in (5.54) is bounded above by $L(\theta_0)$ given in **(H7)**. But the latter is fixed, so we can choose

$$\mathcal{M} = 8CL(\theta_0)/\theta_0 \tag{7.5}$$

independent of $W \in \mathcal{W}$ and still satisfy (5.53) hence (5.52).

Finally, we can close the arguments exactly as we did for Theorem 2.1.

Remark 7.1. Theorems analogous to Theorem 2.2 (respectively, Theorem 2.3) can be proved under hypotheses **(H1)-(H5), (H7)** (respectively, **(H1)–(H6)**).

Examples. **(H7)** holds trivially for

(1) $\mathcal{W} = \{\varepsilon W(t,x) :\ \varepsilon \in \mathbb{R},\ W \text{ fixed}\}$ or

(2) $\mathcal{W} = \{\varepsilon W(\varepsilon^{-1}t, x) : \ \varepsilon \in \mathbb{R} - \{0\}, \ |\varepsilon| \leq 1, \ \ W \text{ fixed}\}$.

In Example (1), ε cancels in the formula $|||W|||^{-2}\sigma(t)$ while in Example (2) we have a time dilation which shrinks the gaps between the almost periods, so the $L(\theta_0)$ valid for $W(t,x)$ is good for the entire family.

(3) There are more general families of perturbations $\mathcal{W}$ for which **(H7)** holds. For example, if $\mathcal{W}$ is equi-almost periodic, see section 9.

8. Appendix: Singular operators. In this section we present the definition and the properties we needed previously for the singular operators

$$e^{-iH_0t}\left(H_0 - \Lambda - i0\right)^{-1}\mathbf{P_c}, \ \delta\left(H_0 - \Lambda\right)\mathbf{P_c}, \ \text{P.V.}\left(H_0 - \Lambda\right)^{-1}\mathbf{P_c}$$

and establish the identities

$$\left(H_0 - \Lambda \mp i0\right)^{-1}\mathbf{P_c} \ = \ \text{P.V.}\left(H_0 - \Lambda\right)^{-1}\mathbf{P_c} \pm i\pi\delta\left(H_0 - \Lambda\right)\mathbf{P_c}$$

suggested by the well-known distributional identities

$$(x \mp i0)^{-1} \ = \ \text{P.V.}\ \frac{1}{x} \ \pm i\pi \ \delta(x).$$

Recall that we are in the complex Hilbert space $\mathcal{H}$ with self-adjoint "weights" $w_\pm$ and projection operator $\mathbf{P_c}$ satisfying (i), (ii), and (iii). We can then construct the complex Hilbert space $\mathcal{H}_+$ as the closure of the domain of w_+ under the scalar product $(f,g)_+ = (w_+f, w_+g)$ and the complex Hilbert space $\mathcal{H}_-$ as the closure of $\mathbf{P_c}\mathcal{H}$ under the scalar product $(f,g)_- = (w_-f, w_-g)$.

By the hypotheses of section 2, H_0 is a self-adjoint operator on $\mathcal{H}$ and satisfies the local decay estimate (2.3). Based on this property, in [11, 22, 23] it is proved that for Λ in the continuous spectrum of H_0 and $t \in \mathbb{R}$

$$T_t \equiv i \lim_{\eta \searrow 0} \int_t^\infty e^{-i(H_0 - \Lambda - i\eta)s} ds \mathbf{P_c},$$
$$T_t^* \equiv -i \lim_{\eta \searrow 0} \int_{-\infty}^{-t} e^{-i(H_0 - \Lambda + i\eta)s} ds \mathbf{P_c}$$

are well defined linear bounded operators from $\mathcal{H}_+$ to $\mathcal{H}_-$. We then define

$$e^{-iH_0t}\left(H_0 - \Lambda - i0\right)^{-1}\mathbf{P_c} \equiv e^{-i\Lambda t}T_t, \tag{8.1}$$
$$e^{+iH_0t}\left(H_0 - \Lambda + i0\right)^{-1}\mathbf{P_c} \equiv e^{+i\Lambda t}T_t^*, \tag{8.2}$$

and

$$\text{P.V.}\left(H_0 - \Lambda\right)^{-1}\mathbf{P_c} \equiv \frac{1}{2}(T_0 + T_0^*), \tag{8.3}$$
$$\delta\left(H_0 - \Lambda\right)\mathbf{P_c} \equiv \frac{1}{2\pi i}(T_0 - T_0^*). \tag{8.4}$$

Note that the definitions imply the identities

$$\left(H_0 - \Lambda \mp i0\right)^{-1}\mathbf{P_c} = \text{P.V.}\left(H_0 - \Lambda\right)^{-1}\mathbf{P_c} \pm i\pi\delta\left(H_0 - \Lambda\right)\mathbf{P_c}. \tag{8.5}$$

Particularly important properties of these operators are their symmetries when viewed as quadratic forms on $\mathcal{H}_+ \times \mathcal{H}_+$. For example, on any $f, \ g \in \mathcal{H}_+$ the quadratic form induced by T_t is given by

$$(f,g) \mapsto (w_+f, w_-T_tg).$$

Note that

$$\lim_{\eta\searrow 0}(f,T_t^\eta g)\equiv\lim_{\eta\searrow 0}\left(f,i\int_t^\infty e^{-i(H_0-\Lambda-i\eta)s}ds\mathbf{P_c}g\right) = (w_+f,w_-T_tg) \tag{8.6}$$

by the following calculation:

$$\begin{aligned}\lim_{\eta\searrow 0}\left(f,i\int_t^\infty e^{-i(H_0-\Lambda-i\eta)s}ds\mathbf{P_c}g\right)&=\lim_{\eta\searrow 0}\left(f,\mathbf{P_c}i\int_t^\infty e^{-i(H_0-\Lambda-i\eta)s}ds\mathbf{P_c}g\right)\\&=\lim_{\eta\searrow 0}\left(f,w_+w_-\mathbf{P_c}i\int_t^\infty e^{-i(H_0-\Lambda-i\eta)s}ds\mathbf{P_c}g\right)\\&=\lim_{\eta\searrow 0}\left(w_+f,w_-i\int_t^\infty e^{-i(H_0-\Lambda-i\eta)s}ds\mathbf{P_c}g\right)\\&=(w_+f,w_-T_tg),\end{aligned}$$

where we used that $\mathbf{P_c}$ is a projection operator commuting with the integral operator, the identity $w_+w_-\mathbf{P_c}=\mathbf{P_c}$ on $\mathcal{H}$, the self-adjointness of $w_\pm$ and $\mathbf{P_c}$, and $\lim_{\eta\searrow 0}w_-T_t^\eta=w_-T_t$ in $\mathcal{L}(\mathcal{H}_+,\mathcal{H})$.

Identity (8.6) suggests the notation

$$(f,g)\mapsto(f,T_tg)$$

for the quadratic form induced by T_t, where $(\cdot,\cdot)$ can formally be treated as the scalar product in $\mathcal{H}$. Moreover, (8.6) implies

$$(f,T_tg)=(T_t^*f,g).$$

Therefore, the quadratic form induced by P.V.$(H_0-\Lambda)^{-1}\mathbf{P_c}$ is the symmetric part of the one induced by T_0 while $\delta(H_0-\lambda)\mathbf{P_c}$ induces the skew-symmetric part of it divided by the factor $i\pi$. As a consequence both the forms corresponding to the last two operators are symmetric.

In conclusion, for any $f,g\in Domain(w_+),\ t\in\mathbb{R}$, and $\Lambda\in\sigma_{\text{cont}}(H_0)$ we have

$$\begin{aligned}(f,e^{\mp iH_0t}(H_0-\Lambda\mp i0)^{-1}\mathbf{P_c}g)&\equiv\big(w_+f,w_-e^{\mp iH_0t}(H_0-\Lambda\mp i0)^{-1}\mathbf{P_c}g\big)\\&\le C_t\ \|w_+f\|\ \|w_+g\|,\end{aligned}\tag{8.7}$$

$$(f,\delta(H_0-\Lambda)\mathbf{P_c}g)\equiv(w_+f,w_-\delta(H_0-\Lambda)\mathbf{P_c}g)\le\ \frac{C_0}{\pi}\ \|w_+f\|\ \|w_+g\|,\tag{8.8}$$

$$(f,\text{P.V.}(H_0-\Lambda)^{-1}\mathbf{P_c}g)\equiv\big(f,w_-\text{P.V.}(H_0-\Lambda)^{-1}\mathbf{P_c}g\big)\le\ C_0\ \|w_+f\|\ \|w_+g\|.\tag{8.9}$$

The inequalities are due to the boundedness of T_t, where C_t denotes the norm of T_t in $\mathcal{L}(\mathcal{H}_+,\mathcal{H}_-)$. Moreover, the following symmetry properties hold:

$$\begin{aligned}(f,e^{\mp iH_0t}(H_0-\Lambda\mp i0)^{-1}\mathbf{P_c}g)&=(e^{\pm iH_0t}(H_0-\Lambda\pm i0)^{-1}\mathbf{P_c}f,g),\\(f,\delta(H_0-\Lambda)\mathbf{P_c}g)&=(\delta(H_0-\Lambda)\mathbf{P_c}f,g),\\(f,\text{P.V.}(H_0-\Lambda)^{-1}\mathbf{P_c}g)&=(\text{P.V.}(H_0-\Lambda)^{-1}\mathbf{P_c}f,g).\end{aligned}$$

9. Appendix: Almost periodic functions. In this section we present the definition and the properties of almost periodic functions we used throughout this paper. We will confine to functions of the form $f:\mathbb{R}\to X$, where X is a complex Banach space with norm denoted by $\|\cdot\|$.

Definition 9.1. *We say that*

$$f : \mathbb{R} \to X$$

is almost periodic if and only if it is continuous and for each $\varepsilon > 0$ there exists a length $L(\varepsilon, f) > 0$ such that in any closed interval of length greater or equal than $L(\varepsilon, f)$ there is at least one τ with the property that for all $t \in \mathbb{R}$ we have

$$\|f(t+\tau) - f(t)\| \leq \varepsilon. \tag{9.1}$$

The number τ with the property above is called an ε almost period for f.

We say that the family, $\mathcal{F}$ of almost periodic functions is equi-almost periodic if $L(\varepsilon, f)$ can be choosen independently of $f \in \mathcal{F}$.

Example. Any continuous periodic function is almost periodic since for any $\varepsilon > 0$ we can choose the length $L(\varepsilon)$ to be the period of the function.

Theorem 9.1. *Any almost periodic function has a relative compact image.*

The proof of the theorem can be found in [9, Property 1, p. 2]. In particular, any almost periodic function $f : \mathbb{R} \to X$ is in the Banach space of all bounded and continuous functions on $\mathbb{R}$ with values in X, $C(X)$, endowed with the uniform norm. The next result is Bochner's characterization of almost periodic functions; see, for example, [9, Bochner's theorem, p. 4].

Theorem 9.2 (Bochner). *Let $f : \mathbb{R} \to X$ be a continuous function. For f to be almost periodic it is necessary and sufficient that the family of functions $\{f(t+h)\}$, $-\infty < h < \infty$, is relatively compact in $C(X)$.*

As a consequence of Bochner's criterion and Property 4 from [9, p. 3] we have the following.

Theorem 9.3. *Suppose $X_1,\ X_2, \ldots, X_{k+1}$ are Banach spaces, $f_i : \mathbb{R} \to X_i,\ 1 \leq i \leq k$ are almost periodic functions, and $g : \prod_{i=1}^{k} \to X_{k+1}$ is continuous. Then $g(f_1(t),\ f_2(t), \ldots, f_k(t))$ is an almost periodic function.*

The last theorem has very important consequences in the theory of almost periodic functions. We will list only those which are useful in our presentation.

Corollary 9.1. *A finite sum of almost periodic functions with values in the same Banach space is an almost periodic function.*

Corollary 9.2. *A product between a complex valued almost periodic function and an arbitrary almost periodic function is an almost periodic function.*

Corollary 9.3. *If $\mathcal{H}$ is a complex Hilbert space, $\mathcal{L}(\mathcal{H})$ is the Banach space of the bounded linear operators on $\mathcal{H}$, and $W : \mathbb{R} \to \mathcal{L}(\mathcal{H})$ is an almost periodic function, then for any $\varphi,\ \psi \in \mathcal{H}$ the following functions are almost periodic:*

$$\begin{aligned}
&t \to W(t)\varphi, \\
&t \to (\psi,\ W(t)\varphi)\,, \\
&t \to (W(t)\psi,\ W(t)\varphi)\,,
\end{aligned}$$

where $(\cdot, \cdot)$ denotes the scalar product on $\mathcal{H}$.

Another essential result in the theory of almost periodic functions is (see, for example, [9, Property 3, p. 3]) the following.

Theorem 9.4. *Any uniform convergent sequence of almost periodic functions converges towards an almost periodic function.*

Corollary 9.4. *If $\{\mu_j\}_{j\in\mathbb{Z}} \subset \mathbb{R}$ and $\{\beta_j\}_{j\in\mathbb{Z}} \subseteq X$ satisfies $\sum_{j\in\mathbb{Z}} \|\beta_j\| < \infty$, then*

$$\sum_{j\in\mathbb{Z}} e^{i\mu_j t}\beta_j$$

is an X-valued almost periodic function of t.

Proof. According to Weierstrass's criterion the series $\sum_{j\in\mathbb{Z}} e^{i\mu_j t}\beta_j$ is uniformly convergent on $\mathbb{R}$.

By Corollary 9.1 and the example above the partial sums of the above series are almost periodic. The result follows now from Theorem 9.4. □

We continue with the harmonic analysis results for almost periodic functions.

THEOREM 9.5 (mean value). *If $f : \mathbb{R} \to X$ is almost periodic, then the following limit exists and it is approached uniformly with respect to $a \in \mathbb{R}$:*

$$\lim_{t\to\infty} \frac{1}{t}\int_a^{a+t} f(s)ds = M(f) \in X.$$

Moreover, whenever

$$t \geq \frac{4\sup_{s\in\mathbb{R}} \|f(s)\| L(\varepsilon/2, f)}{\varepsilon}$$

we have

$$\left\| M(f) - \frac{1}{t}\int_a^{a+t} f(s)ds \right\| \leq \varepsilon$$

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

The proof of the mean value theorem in this form can be found in [2, pp. 39–44]. Note that although Bohr's book considers only complex valued almost periodic functions the proof can be carried on to Banach space valued functions by simply replacing the modulus by the norm and the Lebesque's integral for complex valued functions by the Bochner's integral.

The results of the next theorem are presented in [9, Chapter 2].

THEOREM 9.6 (fundamental theorem). *If $f,\ g : \mathbb{R} \to X$ are almost periodic, then*

(a) *for any $\mu \in \mathbb{R}$,*

$$\lim_{t\to\infty} \frac{1}{t}\int_0^{t} f(s)e^{-i\mu s}ds = a(\mu, f)$$

exists and is nonzero for at most a denumerable set of μ's; if $a(\mu, f) \neq 0$, then $a(\mu, f)$ is called a Fourier coefficient for f while μ is called a Fourier exponent;

(b) *$a(\mu, f) = a(\mu, g)$ for all $\mu \in \mathbb{R}$ if and only if $f \equiv g$;*

(c) *let $\Lambda(f) = \{\mu : \ a(\mu, f) \neq 0\}$ denote the set of Fourier exponents for f; then there is an ordering on $\Lambda(f)$, $\Lambda(f) = \{\mu_1,\ \mu_2, \ldots\}$ independent of the Fourier coefficients, such that for any $\varepsilon > 0$ there exist the numbers $N(\varepsilon) \in \mathbb{N}$, $0 \leq k_{n,\varepsilon} \leq 1$, $n \in \mathbb{N}$, with the property that the trigonometric polynomial*

$$P_\varepsilon(t) = \sum_{n=1}^{N(\varepsilon)} k_{n,\varepsilon} a(\mu_n, f)e^{i\mu_n t}$$

satisfies

$$\|f(t) - P_\varepsilon(t)\| \leq \varepsilon \quad \text{for all } t \in \mathbb{R}.$$

Moreover, $k_{n,\varepsilon}$ can be choosen such that for any fixed n, $\lim_{\varepsilon\searrow 0} k_{n,\varepsilon} = 1$.

In this paper we use a less general result than the above fundamental theorem, namely, the following.

Corollary 9.5. *If $f(t) = \sum_{j\in\mathbb{Z}} e^{i\mu_j t}\beta_j$, where $\{\mu_j\}_{j\in\mathbb{Z}} \subset \mathbb{R}$ and $\sum_{j\in\mathbb{Z}} \|\beta_j\| < \infty$, then $\Lambda(f) = \{\mu_j\}_{j\in\mathbb{Z}}$, $a(\mu_j, f) = \beta_j$, $j \in \mathbb{Z}$, in particular if $\mu_j \neq 0$, $j \in \mathbb{Z}$, then $M(f) = 0$. Moreover, we can arbitrarily order $\Lambda(f)$ and still have that for any $\varepsilon > 0$ there exists a natural number $N(\varepsilon)$ such that*

$$\|f(t) - \sum_{j=-N(\varepsilon)}^{j=N(\varepsilon)} e^{i\mu_j t}\beta_j\| \leq \varepsilon$$

or, in other words, in this particular case the conclusion in part (c) of the fundamental theorem is valid even if we have an arbitrary order on $\Lambda(f)$ and we choose $k_{j,\varepsilon} \equiv 1$.

Proof. By the Weierstrass criterion the series

$$f(t)e^{-i\mu t} = \sum_{j\in\mathbb{Z}} e^{i(\mu_j-\mu)t}\beta_j$$

is uniformly convergent on $\mathbb{R}$. So, when we compute $a(\mu, f)$ we can integrate term by term and therefore use the identities

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t e^{-i\lambda s} ds = \begin{cases} 0 & \text{if} \quad \lambda \neq 0, \\ 1 & \text{if} \quad \lambda = 0 \end{cases}$$

to get the first part of the corollary. The last part is a direct consequence of the fact that f is an absolute and uniform convergent series. □

Acknowledgments. The authors wish to thank A. Soffer and P. D. Miller for discussions on this work.

REFERENCES

[1] V. Bach, J. Fröhlich, I.M. Sigal, and A. Soffer, *Positive commutators and the spectrum of Pauli-Fierz Hamiltonians of atoms and molecules*, Comm. Math. Phys., 207 (1999), pp. 557–587.

[2] H. Bohr, *Almost Periodic Functions*, Chelsea, New York, 1951.

[3] C. Cohen-Tannoudji, J. Dupont-Roc, and G. Grynberg, *Atom-Photon Interactions*, Wiley, New York, 1992.

[4] A. Galindo and P. Pascual, *Quantum Mechanics*, II, Springer, Berlin, 1991.

[5] A. Jensen and T. Kato, *Spectral properties of Schrödinger operators and time-decay of wave functions*, Duke Math. J., 46 (1979), pp. 583–611.

[6] J.-L. Journé, A. Soffer, and C. Sogge, *$L^p \to L^{p'}$ estimates for the time dependent Schrödinger equation*, Bull. Amer. Math. Soc., 23 (1990), pp. 519–524.

[7] E. Kirr and M.I. Weinstein, *Almost Periodic Perturbations of Hamiltonian PDEs with Multiple Bound States*, in preparation.

[8] W. Kohler and G.C. Papanicolaou, *Wave propagation in a randomly inhomogeneous ocean*, in Wave Propagation in Underwater Acoustics, J.B Keller and J.S. Papadakis, eds., Lecture Notes in Phys. 70, Springer, Berlin, 1977.

[9] B.M. Levitan and V.V. Zhibov, *Almost Periodic Functions and Differential Equations*, Cambridge, University Press, London, New York, 1982.

[10] L.D. Landau and E.M. Lifshitz, *Quantum Mechanics: Non-relativistic Theory*, Course of Theoretical Physics, Vol. 3, Pergamon Press, Oxford, UK, 1965.

[11] M. Merkli and I.M. Sigal, *A time-dependent theory of quantum resonances*, Comm. Math. Phys., 201 (1999), pp. 549–576.

[12] P.D. Miller, A. Soffer, and M.I. Weinstein, *Metastability of breather modes of time dependent potentials*, Nonlinearity, 13 (2000), pp. 507–568.

[13] D. Marcuse, *Theory of Dielectric Optical Waveguides*, Academic Press, New York, 1974.

[14] M. Murata, *Rate of decay of local energy and spectral properties of elliptic operators*, Japan J. Math. (N.S.), 6 (1980), pp. 77–127.

[15] R.G. Newton, *Scattering Theory of Waves and Particles*, 2nd ed., Springer-Verlag, New York, 1982.

[16] J. Rauch, *Local decay of scattering solutions to Schrödinger's equation*, Comm. Math. Phys., 61 (1978), pp. 149–168.

[17] M. Reed and B. Simon, *Methods in Modern Mathematical Physics*, I, *Functional Analysis,* Academic Press, New York, 1972.

[18] T. Schonbek, *Decay of solutions of Schrödinger equations*, Duke Math. J., 46 (1979), pp. 203–213.

[19] I.M. Sigal, *Nonlinear wave and Schrödinger equations* I, *Instability of time-periodic and quasiperiodic solutions*, Comm. Math. Phys., 153 (1993), pp. 297–320.

[20] I.M. Sigal, *General characteristics of nonlinear dynamics*, in Spectral and Scattering Theory, M. Ikawa, ed., Proceedings of the Taniguchi International Workshop, Marcel Dekker, Inc., New York, Basel, Hong Kong, 1994, pp. 197–217.

[21] A. Soffer and M.I. Weinstein, *Dynamic theory of quantum resonances and perturbation theory of embedded eigenvalues*, in Partial Differential Equations and Their Applications, P. Greiner, V. Ivrii, L. Seco, and C. Sulem, eds., CRM Proc. Lecture Notes 12, AMS, Providence, 1997, pp. 277–282.

[22] A. Soffer and M.I. Weinstein, *Nonautonomous Hamiltonians*, J. Statist. Phys., 93 (1998), pp. 359–391.

[23] A. Soffer and M.I. Weinstein, *Time dependent resonance theory*, Geom. Funct. Anal., 8 (1998), pp. 1086–1128.

[24] A. Soffer and M.I. Weinstein, *Resonances, radiation damping and instability in Hamiltonian nonlinear wave equations*, Invent. Math., 136 (1999), pp. 9–74.

[25] B. Vainberg, *Scattering of waves in a medium depending periodically on time*, Asterisque, 210 (1992), pp. 327–340.

[26] K. Yajima, *Scattering theory for Schrödinger operators with potentials periodic in time*, J. Math. Soc. Japan, 29 (1977), pp. 729–743.

[27] K. Yajima, *Resonances for the AC-Stark effect*, Comm. Math. Phys., 78 (1982), pp. 331–352.

[28] K. Yajima, *A multichannel scattering theory for some time dependent Hamiltonians, charge transfer problem*, Comm. Math. Phys., 75 (1980), pp. 153–178.

SIAM J. Math. Anal.
Vol. 33, No. 1, pp. 53–83

HOMOGENIZATION OF ELLIPTIC DIFFERENCE OPERATORS[*]

ANDREY PIATNITSKI[†] AND ELISABETH REMY[‡]

Abstract. We develop some aspects of general homogenization theory for second order elliptic difference operators and consider several models of homogenization problems for random discrete elliptic operators with rapidly oscillating coefficients. More precisely, we study the asymptotic behavior of effective coefficients for a family of random difference schemes whose coefficients can be obtained by the discretization of random high-contrast checker-board structures. Then we compare, for various discretization methods, the effective coefficients obtained with the homogenized coefficients for corresponding differential operators.

Key words. random media, homogenization, H-convergence, difference operator, percolation, random walk

AMS subject classifications. 35B40, 39A10

PII. S003614100033808X

1. Introduction.

We develop some aspects of general H-convergence and homogenization theory for second order elliptic difference operators and consider several homogenization problems for random discrete elliptic operators with rapidly oscillating coefficients. More precisely, we study the asymptotic behavior of effective coefficients for a family of random difference schemes whose coefficients can be obtained by the discretization of random high-contrast checker-board structures. Then we compare, for various discretization methods, the effective coefficients obtained with the homogenized coefficients for corresponding differential operators.

Many results can also be formulated in terms of the central limit theorem for random walks in random statistically homogeneous media.

Originally, G- and H-convergence of differential operators and Γ-convergence of the corresponding functionals were introduced by Spagnolo [27], De Giorgi [7], [8], and Murat and Tartar [22]. Then these notions were developed and generalized essentially in the works of Bensoussan, Lions, and Papanicolaou [4], Tartar [26], Murat [21], Jikov et al. [28], G. Dal Maso [18], and many others. This resulted in the appearance of advanced homogenization theory.

In recent years, significant progress has been achieved in the homogenization theory of random differential operators. We refer to the original works of Kozlov [13] and Papanicolaou and Varadhan [24], and to the book by Jikov, Kozlov, and Oleinik [11] wherein an additional bibliography can be found. In particular, in case of random high-contrast checker-board structures, the asymptotics of effective diffusion have been constructed in Jikov, Kozlov, and Oleinik [11]. Berlyand and Golden in [5] have improved this result in a special case.

In contrast with differential operators, the homogenization theory of difference operators is not so well developed. There are only a few mathematical works on this subject, among them Künnemann [17], Kozlov [14], [15], and Krasniansky [16]. In

[*] Received by the editors January 20, 2000; accepted for publication (in revised form) December 5, 2000; published electronically April 12, 2001.

http://www.siam.org/journals/sima/33-1/33808.html

[†] P. N. Lebedev Physical Institute RAS, Leninski prospect 53, Moscow 117924, Russia (andrey@sci.lpi.msk.su).

[‡] INRIA/LATP, 38 rue Joliot–Curie, 13451 Marseille Cedex 13, France (eremy@sophia.inria.fr). Current address: Laboratoire “Genome et Informatique,” Université de Versailles-St. Quentin, Batiment Buffon, 45, avenue des Etats-Unis, 78035 Versailles, France (remy@genetique.uvsq.fr).

[17] it is proved that the central limit theorem holds for symmetric random walks in random ergodic statistically homogeneous media. Then, many interesting results for various kinds of random walks in random media were obtained in Kozlov [14]. The first homogenization results for difference schemes were formulated and proved in Kozlov [15]. We also mention the work Bricmont and Kupiainen [6] where the central limit theorem was obtained for a class of nonsymmetric random walks.

Perhaps the difference operators with rapidly oscillating coefficients did not attract the attention of mathematicians because these operators did not appear in the classical difference schemes approximation approach (see, for example, Quarteroni and Valli [25]): the fast oscillation of coefficients of difference schemes would contradict the regularity and even the measurability of coefficients of the initial differential equations.

On the other hand, many modern practical and numerical applications involve various homogenization problems for discrete operators with rapidly oscillating coefficients. For instance, when discretizing microinhomogeneous media, due to the natural restrictions, it is not possible to keep the size of the numerical grid much smaller than the typical size of inhomogeneity (the microscopic length scale) of the medium. This leads to the appearance of difference operators with rapidly oscillating coefficients (see, for instance, McCarthy [19], Nœtinger [23]). The most important question here is, How far could the effective coefficients of a difference scheme diverge from ones of corresponding differential operators? The first successful attempt to answer this question was done by Avellaneda, Hou, and Papanicolaou [2] where it was shown that, in the multidimensional case, the finite difference approach does not provide the right homogenized coefficients unless the ratio of the size of a discretization mesh to the microscopic length scale goes to 0.

In the present work we show that the effective coefficients of the difference schemes approximating a family of elliptic PDEs with rapidly oscillating coefficients depend essentially on the discretization method.

The paper is divided into two parts. The first one is devoted to H-convergence and homogenization of difference operators.

Earlier homogenization problems for difference operators were investigated by Kozlov in [15] where a number of homogenization results for difference schemes were obtained. In the present work we extend further the homogenization theory of discrete operators and prove a number of basic statements such as convergence of solutions of the Neumann problem, convergence of energies and of arbitrary solutions, Γ-convergence, and some others. To this end we mainly use the discrete analogue of the compensated compactness technique originally introduced in Murat [21] and Tartar [26] for functions of continuous arguments. Namely, we prove a version of compensated compactness lemma, adapted to difference operators, and then apply it systematically in our considerations in combination with the method of correctors and variational techniques.

For the sake of completeness we also formulate some technical results from Kozlov [15] and give another proof of the homogenization theorem for random difference operators. An additional reason for this is the fact that we use a more general definition of ellipticity than that in [15].

It should be noted that although some basic ideas here have been borrowed from homogenization theory of differential equations, still the peculiarities of difference operators such as the big dimension of difference gradient, the irreducibility and ellipticity conditions in the case of boundary-value problems, and the asymptotic nature of difference schemes, create additional difficulties in studying these operators and make the generalization of homogenization theory to difference operators nontrivial.

In the second part of the paper, we discretize high-contrast two-dimensional checker-board structures, find the asymptotics of effective diffusion, and show that different discretization methods lead to different asymptotics.

1.1. Difference elliptic operators. Let $Q \subset \mathbb{R}^d$ be a smooth bounded domain and let $Q_\varepsilon = Q \cap \varepsilon\mathbb{Z}^d$, where $\mathbb{Z}^d$ is the standard integer lattice in $\mathbb{R}^d$ and $\varepsilon > 0$. We consider the discrete Dirichlet problem in Q_ε:

$$A_\varepsilon u^\varepsilon(x) = \sum_{z,z' \in \Lambda} \partial^\varepsilon_{-z}\left(a^\varepsilon_{zz'}(x)\partial^\varepsilon_{z'} u^\varepsilon(x)\right) = f^\varepsilon(x) \quad \text{in } Q_\varepsilon\,, \qquad u^\varepsilon(x) = 0 \quad \text{on } \partial Q^\Lambda_\varepsilon\,. \tag{1.1}$$

Here Λ is a fixed finite subset of $\mathbb{Z}^d$ symmetric with respect to 0, the matrix $\mathcal{A}^\varepsilon = \{a^\varepsilon_{zz'}\}$ is symmetric, $\partial Q^\Lambda_\varepsilon$ is the boundary of Q_ε defined by

$$\partial Q^\Lambda_\varepsilon \overset{\triangle}{=} (Q_\varepsilon + \varepsilon\Lambda) \setminus Q_\varepsilon = \{x + \varepsilon z \,|\, x \in Q_\varepsilon,\, z \in \Lambda\} \setminus Q_\varepsilon\,,$$

and ∂^ε_z is the standard difference derivative: $(\partial^\varepsilon_z v)(x) \overset{\triangle}{=} \frac{1}{\varepsilon}\left(v(x + \varepsilon\, z) - v(x)\right)$. For any $v^\varepsilon : Q_\varepsilon \mapsto \mathbb{R}$, we introduce the following norm (the $L^2(Q_\varepsilon)$-norm): $\|v^\varepsilon\|^2_{L^2(Q_\varepsilon)} \overset{\triangle}{=} \varepsilon^d \sum_{x \in Q_\varepsilon} |v^\varepsilon(x)|^2$. We say that a function v^ε defined on $\varepsilon\mathbb{Z}^d$ belongs to the space $W^{1,2}_0(Q_\varepsilon)$ if $v(x) = 0$ for $x \notin Q_\varepsilon$. We define the norm on the space $W^{1,2}_0(Q_\varepsilon)$ as follows: $\|v^\varepsilon\|^2_{W^{1,2}_0(Q_\varepsilon)} = \varepsilon^d \sum_{x \in \overline{Q_\varepsilon}} \sum_{i=1}^d |\partial^\varepsilon_{\pm e_i} v^\varepsilon(x)|^2$, where $\{e_i\}_{i=1,\ldots,d}$ is the standard basis in $\mathbb{R}^d$ and $\overline{Q_\varepsilon} \overset{\triangle}{=} Q_\varepsilon + \varepsilon\Lambda = Q_\varepsilon \cup \partial Q^\Lambda_\varepsilon$; $W^{-1,2}(Q_\varepsilon)$ is the dual space to $W^{1,2}_0(Q_\varepsilon)$.

In the summation in (1.1), we can consider only the elements from the set $\Lambda \setminus \{0\}$, as the contribution of the element $\{0\}$ is null.

Definition 1.1. *We say that the family of problems* (1.1) *(or, simply, problem* (1.1)*) is* uniformly elliptic *if there are* $c_1, c_2 > 0$ *and* $\varepsilon_0 > 0$ *such that, for any* $v^\varepsilon \in W^{1,2}_0(Q_\varepsilon)$ *and any* $\varepsilon < \varepsilon_0$,

$$|a^\varepsilon_{zz'}(x)| \le c_1, \tag{1.2}$$

$$c_2\|v^\varepsilon\|^2_{W^{1,2}_0(Q_\varepsilon)} \le \varepsilon^d \sum_{x \in \overline{Q_\varepsilon}} \sum_{z,z' \in \Lambda} a^\varepsilon_{zz'}(x)\partial_{z'} v^\varepsilon(x)\partial_z v^\varepsilon(x)\,. \tag{1.3}$$

Remark 1.2. *The uniform boundedness of the matrix* $\mathcal{A}^\varepsilon$ *implies the following upper bound:*

$$\varepsilon^d \sum_{x \in \overline{Q_\varepsilon}} \sum_{z,z' \in \Lambda} a^\varepsilon_{zz'}(x)\partial_{z'} v^\varepsilon(x)\partial_z v^\varepsilon(x) \le c(\Lambda)\|v^\varepsilon\|^2_{W^{1,2}_0(Q_\varepsilon)}\,.$$

Indeed, it suffices to represent z *as a sum* $z = z^1 + z^2 + \cdots + z^N$ *with* $|z^i| = 1$ *for all* $i = 1 \ldots N$. *Then,*

$$\partial^\varepsilon_z v^\varepsilon(x) = \sum_{k=1}^N \partial^\varepsilon_{z^k} v^\varepsilon(x + z^1 + \cdots + z^{k-1})$$

and the required bound is the consequence of the finiteness of Λ.

In what follows we always assume the uniform ellipticity conditions (1.2)–(1.3) to hold.

It should be noted that, in general, the uniform ellipticity condition (1.3) is rather implicit. For instance, it neither requires the positiveness of the matrix $\{a^{\varepsilon}_{zz'}(x)\}$ nor follows from the estimate

$$c_3|\xi|^2 \le \sum_{z,z'\in\Lambda} a^{\varepsilon}_{zz'}(x)(\xi,z)(\xi,z') \le c_4|\xi|^2, \quad \xi\in\mathbb{R}^d,\ c_3>0\,, \tag{1.4}$$

where $(\cdot,\cdot)$ is the scalar product in $\mathbb{R}^d$. One can easily see this by considering the one-dimensional problem with

$$a^{\varepsilon}_{zz'}(x) = \begin{cases} 1/2 & \text{if } z=z',\ |z|=2,\\ 0 & \text{otherwise.}\end{cases}$$

Clearly, (1.3) is not satisfied although (1.4) holds.

In order to ensure the uniform ellipticity of problem (1.1) one should combine estimates such as (1.4) with a proper irreducibility condition. Below we show that for two important particular classes of difference operators commonly used in applications, the ellipticity conditions can be easily verified.

Suppose we are given a family of functions $p^{\varepsilon}_z(x)$, $x\in Q_\varepsilon$, $z\in\Lambda$, possessing the following properties:

1. positiveness: $p^{\varepsilon}_z(x)\ge 0,\ \sum_{z\in\Lambda}p^{\varepsilon}_z(x)=1$ for each $x\in Q_\varepsilon$,
2. $p^{\varepsilon}_{\pm e_i}(x)\ge\delta>0$, $i=1,\dots,d$,
3. symmetry: $p^{\varepsilon}_z(x)=p^{\varepsilon}_{-z}(x+\varepsilon z)$.

Then, the family of problems

$$u^\varepsilon(x)=\sum_{z\in\Lambda}p^{\varepsilon}_z(x)\,u^\varepsilon(x+\varepsilon z)+\varepsilon^2 f^\varepsilon(x)\quad\text{in } Q_\varepsilon\,,\qquad u^\varepsilon(x)=0\quad\text{on } \partial Q^{\Lambda}_{\varepsilon}\,, \tag{1.5}$$

can be easily rewritten in the form (1.1) with

$$a^{\varepsilon}_{zz'}(x) = \begin{cases} p^{\varepsilon}_z(x) & \text{if } z=z',\ z\neq 0,\\ 0 & \text{otherwise.}\end{cases} \tag{1.6}$$

Proposition 1.3. *Let* $\{p^{\varepsilon}_z(x)\}$ *possess the abovementioned properties* (1), (2), *and* (3). *Then problem* (1.5) *is uniformly elliptic.*

Proof. Summing by parts, one can show after simple calculations that

$$\begin{aligned}\delta\varepsilon^{-d}\|v^\varepsilon\|^2_{W^{1,2}_0(Q_\varepsilon)} &= \delta\sum_{x\in\overline{Q_\varepsilon}}\sum_{i=1}^{d}|\partial^{\varepsilon}_{\pm e_i}v^\varepsilon|^2\\ &\le \sum_{x\in\overline{Q_\varepsilon}}\sum_{z,z'\in\Lambda}a^{\varepsilon}_{zz'}(x)\partial_{z'}v^\varepsilon(x)\partial_z v^\varepsilon(x)\\ &\le C\sum_{x\in\overline{Q_\varepsilon}}\sum_{z\in\Lambda}|\partial^{\varepsilon}_z v^\varepsilon|^2 \le c(\Lambda)\varepsilon^{-d}\|v^\varepsilon\|^2_{W^{1,2}_0(Q_\varepsilon)}\end{aligned}$$

uniformly in ε. This yields the desired result. □

Assumption (2) can be relaxed as follows:

(2′) For some $N > 0$ and $\delta > 0$ and for any $x', x'' \in Q_\varepsilon$, $|x' - x''| = \varepsilon$, there is a finite sequence of vectors $y^1, y^2, \dots, y^k \in \Lambda$, $k \le N$, such that $x'' = x' + \varepsilon \sum_{j=1}^{k} y^j$ and $p^\varepsilon_{y^j}(x' + \varepsilon \sum_{i=1}^{j-1} y^i) \ge \delta$.

Another important class of uniformly elliptic operators is formed by matrices $\{a^\varepsilon_{zz'}(x)\}$ that satisfy the estimate

$$\sum_{z,z' \in \Lambda} a^\varepsilon_{zz'}(x)\, \eta_z\, \eta_{z'} \ge c \sum_{i=1}^{d} |\eta_{\pm e_i}|^2, \qquad \eta \in \mathbb{R}^{|\Lambda|},$$

uniformly in ε and $x \in Q_\varepsilon$; here we assume that all the vectors $\pm e_i$, $i = 1, 2, \dots, d$, are elements of Λ; if it is not the case, the right-hand side (RHS) of the latter formula does not make sense.

Clearly, the uniform ellipticity implies the coerciveness of problem (1.1) and we have the following statement.

Proposition 1.4. *Let problem* (1.1) *be uniformly elliptic and* $f^\varepsilon \in L^2(Q_\varepsilon)$. *Then there exists a unique solution* $u^\varepsilon \in W_0^{1,2}(Q_\varepsilon)$ *and the estimate*

$$\|u^\varepsilon\|_{W_0^{1,2}(Q_\varepsilon)} \le c\|f^\varepsilon\|_{L^2(Q_\varepsilon)}$$

holds uniformly in ε. Henceforth we usually suppose that $f^\varepsilon(\cdot)$ is a discretization of a given function $f \in L^2(Q)$.

We also define the norm on the space $W^{1,2}(Q_\varepsilon)$ by

$$\|v^\varepsilon\|^2_{W^{1,2}(Q_\varepsilon)} = \varepsilon^d \sum_{x \in \overline{Q_\varepsilon}} \sum_{i=1}^{d} |\bar\partial^\varepsilon_{\pm e_i} v^\varepsilon(x)|^2 + \|v^\varepsilon\|^2_{L^2(Q^\varepsilon)},$$

where we use the notation

$$\bar\partial^\varepsilon_z \varphi(x) = \begin{cases} \partial^\varepsilon_z \varphi(x) & \text{if } x + \varepsilon z \in \overline{Q_\varepsilon}, \\ 0 & \text{otherwise.} \end{cases}$$

2. Tools for discrete operators analysis.

2.1. Compensated compactness lemma. One of the main tools in the homogenization of differential operators is the so-called compensated compactness lemma (see Murat [21] and Tartar [26]), which gives a sufficient condition for passing to the limit in the inner product of two weakly converging sequences of vector functions. In this section, we prove the discrete version of this result that serves the case of functions defined on a grid.

First of all, we introduce the discrete divergence as follows: for any vector function $q \in \big(L^2(Q_\varepsilon)\big)^{|\Lambda|}$,

$$\operatorname{div}^\varepsilon_\Lambda q(x) \overset{\triangle}{=} \sum_{z \in \Lambda} \partial^\varepsilon_{-z} q_z(x).$$

It should be emphasized that the above divergence operator depends on the choice of the set Λ.

Lemma 2.1. *Let* q^ε *and* v^ε *be sequences of vector functions from* $\big(L^2(Q_\varepsilon)\big)^{|\Lambda|}$ *such that*

$$q^\varepsilon \xrightarrow[\varepsilon \to 0]{} q^0 \text{ weakly in } L^2(Q_\varepsilon), \operatorname{div}^\varepsilon_\Lambda q^\varepsilon \xrightarrow[\varepsilon \to 0]{} f^0 \text{ in } W^{-1,2}(Q_\varepsilon),$$

$$v^\varepsilon \xrightarrow[\varepsilon \to 0]{} v^0 \text{ weakly in } L^2(Q_\varepsilon), v^\varepsilon_z(x) = \partial^\varepsilon_z u^\varepsilon(x) \text{ for some } u^\varepsilon \in W^{1,2}(Q_\varepsilon).$$

Then, the sequence $(q^\varepsilon\, v^\varepsilon)$ *converges* $\star$*-weakly to* $q^0\, v^0$*:* $\quad q^\varepsilon\, v^\varepsilon \xrightarrow[\varepsilon\to 0]{\star} q^0\, v^0\,.$

Proof. According to Kozlov [15, Proposition 3], the weak convergence of q^ε in $L^2(Q_\varepsilon)$ implies the following weak convergence in $W^{-1,2}(Q_\varepsilon)$:

$$\mathrm{div}_\Lambda^\varepsilon q^\varepsilon \xrightarrow[\varepsilon\to 0]{} \sum_{z\in\Lambda} \frac{\partial}{\partial z} q_z^0 = \sum_{z\in\Lambda} z\,\cdot\,\nabla q_z^0;$$

here the standard notation $\frac{\partial}{\partial z} f(x) = z\,\cdot\,\nabla_x f(x)$ for the derivative along arbitrary vector z has been used. Thus, $\sum_{z\in\Lambda} z\,\cdot\,\nabla q_z^0 = f^0$, and we have

$$\lim_{\varepsilon\to 0} \|\mathrm{div}_\Lambda^\varepsilon (q^\varepsilon - q^0)\|_{W^{-1,2}(Q_\varepsilon)} = 0\,.$$

From now on, the notation like q^0 or v^0 is used both for the functions of continuous argument and for their discretization (see Appendix A). Using the representation $q^\varepsilon\, v^\varepsilon = (q^\varepsilon - q^0)\, v^\varepsilon + q^0\, v^\varepsilon$ and taking into account the $\star$-weak convergence of $q^\varepsilon\, v^0$ to $q^0\, v^0$, one can assume, without loss of generality, that $q^0 = 0$. Also, under the proper choice of additive constant, $\sum_{x\in Q_\varepsilon} u^\varepsilon(x) = 0$. Then, by the Poincaré inequality, the sequence u^ε is uniformly bounded in the $W^{1,2}$-norm. For any $\varphi \in \mathcal{C}_0^\infty(Q)$ we get

$$\begin{aligned}\varepsilon^d \sum_{x\in Q_\varepsilon} q^\varepsilon(x)\, v^\varepsilon(x)\, \varphi(x) \;&= \varepsilon^d \sum_{x\in Q_\varepsilon} \sum_{z\in\Lambda} q_z^\varepsilon(x)\, \partial_z^\varepsilon u^\varepsilon(x)\, \varphi(x) \\ &= \varepsilon^d \sum_{x\in Q_\varepsilon} \sum_{z\in\Lambda} \{q_z^\varepsilon(x)\, \partial_z^\varepsilon(u^\varepsilon(x)\, \varphi(x)) - q_z^\varepsilon(x)\, u^\varepsilon(x)\, \partial_z^\varepsilon \varphi(x)\} + \tau(\varepsilon)\end{aligned}$$

with $\lim_{\varepsilon\to 0} \tau(\varepsilon) = 0$ (see Appendix B). Summing by parts in the latter expression leads to

$$\begin{aligned}&\varepsilon^d \sum_{x\in Q_\varepsilon} q^\varepsilon(x)\, v^\varepsilon(x)\, \varphi(x) \\ &\quad= \varepsilon^d \sum_{x\in Q_\varepsilon} \sum_{z\in\Lambda} \left\{\partial_{-z}^\varepsilon q_z^\varepsilon(x)\, u^\varepsilon(x)\, \varphi(x) - q_z^\varepsilon(x)\, u^\varepsilon(x)\, \partial_z^\varepsilon \varphi(x)\right\} + \tau(\varepsilon) \\ &\quad= \varepsilon^d \sum_{x\in Q_\varepsilon} (\mathrm{div}_\Lambda^\varepsilon\, q^\varepsilon(x)\ , u^\varepsilon\, \varphi) - \varepsilon^d \sum_{x\in Q_\varepsilon} \sum_{z\in\Lambda} q_z^\varepsilon(x)\, u^\varepsilon(x)\, \partial_z^\varepsilon \varphi(x) + \tau(\varepsilon)\,.\end{aligned}$$

Since u^ε is uniformly bounded in $W^{1,2}(Q_\varepsilon)$ and $\mathrm{div}_\Lambda^\varepsilon\, q^\varepsilon$ converges to 0 in the $W^{-1,2}$-norm, the first term in the RHS goes to 0 as $\varepsilon \to 0$. The second term goes to 0 because $q_z^\varepsilon\, \partial_z^\varepsilon \varphi$ converges to 0 in $L^2(Q_\varepsilon)$ weakly. Finally, for any $\varphi \in \mathcal{C}_0^\infty(Q)$, $\lim_{\varepsilon\to 0} \sum_{x\in Q_\varepsilon} \sum_{z\in\Lambda} q_z^\varepsilon(x)\, v_z^\varepsilon(x)\, \varphi(x) = 0\,.$ □

2.2. *H*-convergence and homogenization. In this section, we give the definitions of the H-convergence and the homogenization of discrete operators and then study the main properties of this convergence (see Spagnolo [27], Murat and Tartar [22] for the relevant definitions in case of differential operators).

Consider a family of uniformly elliptic discrete Dirichlet problems,

$$A_\varepsilon u^\varepsilon = \mathrm{div}_\Lambda^\varepsilon \left(\sum_{z'\in\Lambda} a_{zz'}^\varepsilon\, \partial_{z'}^\varepsilon u^\varepsilon\right) = f^\varepsilon\,, \quad u^\varepsilon \in W_0^{1,2}(Q_\varepsilon)\,, \tag{2.1}$$

and denote by $\mathcal{A}^\varepsilon(x)$ the matrices of the coefficients $\{a_{zz'}^\varepsilon(x)\}$. Let $\mathcal{A}(x) = \{a_{zz'}(x)\}$, $x \in Q$, be a $|\Lambda| \times |\Lambda|$ matrix.

Definition 2.2 (H-convergence). *We say that the matrix* $\mathcal{A}^\varepsilon$ H-converges *to* $\mathcal{A}$ *(*$\mathcal{A}^\varepsilon \xrightarrow[\varepsilon\to 0]{H} \mathcal{A}$*) if, for any sequence* $f^\varepsilon \in W^{-1,2}(Q_\varepsilon)$ *such that* $f^\varepsilon \xrightarrow[\varepsilon\to 0]{} f$ *in* $W^{-1,2}(Q_\varepsilon)$*, we have*

$$\begin{aligned} u^\varepsilon \quad &\xrightarrow[\varepsilon\to 0]{} \quad u^0 && \textit{weakly in } W_0^{1,2}(Q_\varepsilon)\,,\\ s^\varepsilon = \sum_{z\in\Lambda} a^\varepsilon_{zz'}\,\partial^\varepsilon_z u^\varepsilon \quad &\xrightarrow[\varepsilon\to 0]{} \quad s^0 = \sum_{z\in\Lambda} a_{zz'}\,\frac{\partial}{\partial z}u^0 && \textit{weakly in } L^2(Q_\varepsilon)\,, \end{aligned}$$

where u^0 *is the solution of the limit Dirichlet problem,*

$$\sum_{z,z'\in\Lambda} -\frac{\partial}{\partial z}\left(a_{zz'}(x)\,\frac{\partial}{\partial z'}u^0\right) = f, \quad u^0 \in W_0^{1,2}(Q).$$

The homogenization is a particular case of H-convergence. Given a matrix-valued function $\mathcal{A}^1(x) = \{a^1_{zz'}(x)\}$, z, $z' \in \Lambda$, $x \in \mathbb{Z}^d$, we define the sequence $\mathcal{A}^\varepsilon$ as follows: $\mathcal{A}^\varepsilon(x) = \mathcal{A}^1(x/\varepsilon)$, $x \in Q_\varepsilon$. Suppose that the corresponding family of problems (defined in (2.1)) is uniformly elliptic.

Definition 2.3. *The constant matrix* $\mathcal{A}$ *is the* homogenized matrix *for* $\mathcal{A}^\varepsilon(x) = \{a^\varepsilon_{zz'}(x)\}$ *if, for any sequence* $f^\varepsilon \in W^{-1,2}(Q_\varepsilon)$ *such that* $f^\varepsilon \xrightarrow[\varepsilon\to 0]{} f$ *in* $W^{-1,2}(Q)$*, the solutions* u^ε *of the Dirichlet problems*

$$\operatorname{div}^\varepsilon_\Lambda\left(\sum_{z'\in\Lambda} a^\varepsilon_{zz'}\,\partial^\varepsilon_{z'}u^\varepsilon\right) = f^\varepsilon,\ \ u^\varepsilon \in W_0^{1,2}(Q_\varepsilon)\,,$$

converge to the solution u^0 *of the limit Dirichlet problem*

$$-\sum_{z,z'\in\Lambda}\frac{\partial}{\partial z}a_{zz'}\frac{\partial}{\partial z'}u^0 = f\,,\ \ u^0 \in W_0^{1,2}(Q), \tag{2.2}$$

in the following sense:

$$\begin{aligned} u^\varepsilon \quad &\xrightarrow[\varepsilon\to 0]{} \quad u^0 && \textit{weakly in } W_0^{1,2}(Q)\,,\\ \sum_{z'\in\Lambda} a^\varepsilon_{zz'}\,\partial^\varepsilon_{z'}u^\varepsilon \quad &\xrightarrow[\varepsilon\to 0]{} \quad \sum_{z\in\Lambda} a_{zz'}\,\frac{\partial}{\partial z}u^0 && \textit{weakly in } L^2(Q)\,. \end{aligned}$$

Remark 2.4. *The dimension of the difference gradient of functions defined on* Q_ε *is equal to* $|\Lambda|$ *and does not coincide with the dimension of the standard gradient of functions defined on* Q*. This is the reason we write the limit equation in the definitions above in a nonstandard form. This allows us to define the convergence of streams. Of course, one can easily transform the limiting equation to the standard form*

$$\sum_{z,z'\in\Lambda}\frac{\partial}{\partial z}a_{zz'}(x)\frac{\partial}{\partial z'} = \sum_{i,j=1}^{d}\frac{\partial}{\partial x_i}\check{a}_{ij}(x)\frac{\partial}{\partial x_j}\,, \quad \check{a}_{ij}(x) = \sum_{z,z'\in\Lambda}(z,e_i)a_{zz'}(x)(z',e_j)\,.$$

One of the remarkable properties of H- and G-convergences of differential operators is the compactness of a family of uniformly elliptic operators; see, for example, Murat and Tartar [22], Zhikov et al. [28]. We proceed by quoting the compactness result for a family of uniformly elliptic difference operators.

Proposition 2.5 (see Kozlov [15, section 2]). *Any uniformly elliptic sequence of problems defined in* (2.1) *contains an H-convergent subsequence. The limit problem involves a second order uniformly elliptic operator in divergence form:*

$$A\,u = -\sum_{z,z'\in\Lambda}\frac{\partial}{\partial z}\left(a_{zz'}(x)\,\frac{\partial}{\partial z'}\,u\right) = -\sum_{i,j=1}^{d}\frac{\partial}{\partial x_i}\left(a_{ij}(x)\,\frac{\partial}{\partial x_j}u\right).$$

In the subsections below we prove a number of general results on H-convergence and homogenization of difference operators that are not exhibited in the existing literature.

2.2.1. Convergence of arbitrary solutions. One of the significant properties of H-convergence is the fact that the H-limit operator depends only on the original sequence of operators and does not depend on the type of boundary conditions and on the domain. In a general form, this can be formulated as follows.

Theorem 2.6 (convergence of arbitrary solutions). *Let a sequence of uniformly elliptic operators A_ε H-converge in a domain Q to the limit operator A, and suppose that a sequence of functions $w^\varepsilon\in W^{1,2}(Q_\varepsilon)$ satisfies the conditions*

$$\begin{aligned} &w^\varepsilon \xrightarrow[\varepsilon\to 0]{} w^0 \quad \textit{weakly in } W^{1,2}(Q_\varepsilon)\,,\\ &\operatorname{div}^\varepsilon_\Lambda\left(\sum_{z'\in\Lambda} a^\varepsilon_{zz'}\,\left(g_{z'}+\partial^\varepsilon_{z'}w^\varepsilon\right)\right) = f\,, \end{aligned} \tag{2.3}$$

where $g\in\left(L^2(Q)\right)^{|\Lambda|}$ and $f\in W^{-1,2}(Q)$ do not depend on ε. Then, w^0 satisfies the homogenized equation

$$-\sum_{z,z'\in\Lambda}\frac{\partial}{\partial z}\left[a_{zz'}\left(g_{z'}+\frac{\partial}{\partial z'}w^0\right)\right] = f\,,$$

and the streams do converge in $L^2(Q_\varepsilon)$ weakly:

$$\sum_{z'\in\Lambda} a^\varepsilon_{zz'}\,\left(g_{z'}+\partial^\varepsilon_{z'}w^\varepsilon\right) \xrightarrow[\varepsilon\to 0]{} \sum_{z'\in\Lambda} a_{zz'}\left(g_{z'}+\frac{\partial}{\partial z'}w^0\right).$$

Proof. Under the conditions of the theorem, the streams are uniformly bounded in $L^2(Q_\varepsilon)$. Thus, taking a proper subsequence, we have $\sum_{z'\in\Lambda} a^\varepsilon_{zz'}\,\left(g_{z'}+\partial^\varepsilon_{z'}w^\varepsilon\right) \xrightarrow[\varepsilon\to 0]{}$ $\xi_z$weakly in $L^2(Q_\varepsilon)$. Passing to the limit in (2.3), one can easily check that $-\sum_{z\in\Lambda}\frac{\partial}{\partial z}\xi_z$ $= f$. We have to prove the relation $\xi_z = \sum_{z'\in\Lambda} a_{zz'}\left(g_{z'}+\frac{\partial}{\partial z'}w^0\right)$. Let u^0 be an arbitrary function from $W^{1,2}_0(Q)$. Denote by u^ε the solution of the Dirichlet problem,

$$\operatorname{div}^\varepsilon_\Lambda\left(\sum_{z'\in\Lambda} a^\varepsilon_{zz'}\,\partial^\varepsilon_{z'}u^\varepsilon\right) = \sum_{z,z'\in\Lambda}\frac{\partial}{\partial z}\left(a_{zz'}\,\frac{\partial}{\partial z'}\,u^0\right),$$

and consider the following identity:

$$\sum_{z\in\Lambda}(g_z+\partial^\varepsilon_z w^\varepsilon)\sum_{z'\in\Lambda} a^\varepsilon_{zz'}\,\partial^\varepsilon_{z'}u^\varepsilon = \sum_{z\in\Lambda}\partial^\varepsilon_z u^\varepsilon\sum_{z'\in\Lambda} a^\varepsilon_{zz'}\,(g_{z'}+\partial^\varepsilon_{z'}w^\varepsilon)\,. \tag{2.4}$$

By the definition of H-convergence, we have

$$\sum_{z'\in\Lambda} a^\varepsilon_{zz'}\,\partial^\varepsilon_{z'} u^\varepsilon \xrightarrow[\varepsilon\to 0]{} \sum_{z'\in\Lambda} a_{zz'}\frac{\partial}{\partial z'}u^0 \quad \text{weakly in } L^2(Q_\varepsilon)\,,$$

while the limiting relation

$$\sum_{z\in\Lambda}(g_z+\partial^\varepsilon_z w^\varepsilon) \xrightarrow[\varepsilon\to 0]{} \sum_{z\in\Lambda}\left(g_z+\frac{\partial}{\partial z}w^0\right) \quad \text{weakly in } L^2(Q_\varepsilon)$$

is an evident consequence of the weak convergence of w^ε . Now, passing to the limit on the left-hand side (LHS) of (2.4), with the help of Lemma 2.1 we obtain

$$\sum_{z\in\Lambda}(g_z+\partial^\varepsilon_z w^\varepsilon)\sum_{z'\in\Lambda} a^\varepsilon_{zz'}\,\partial^\varepsilon_{z'} u^\varepsilon \xrightarrow[\varepsilon\to 0]{\star} \sum_{z\in\Lambda}\left(g_z+\frac{\partial}{\partial z}w^0\right)\sum_{z'\in\Lambda} a_{zz'}\frac{\partial}{\partial z'}u^0\,.$$

The fact that g_z does not depend on ε has also been used here.

Similarly, passing to the limit on the RHS of (2.4) gives

$$\sum_{z\in\Lambda}\partial^\varepsilon_z u^\varepsilon\sum_{z'\in\Lambda} a^\varepsilon_{zz'}\,(g_{z'}+\partial^\varepsilon_{z'} w^\varepsilon) \xrightarrow[\varepsilon\to 0]{\star} \sum_{z\in\Lambda}\frac{\partial}{\partial z}u^0\,\xi_z\,.$$

Finally, considering the fact that u^0 is arbitrary function from $W^{1,2}_0(Q)$, we deduce

$$\xi_z=\sum_{z'\in\Lambda} a^\varepsilon_{zz'}\,\left(g_{z'}+\frac{\partial}{\partial z'}w^0\right)\,. \qquad \square$$

Corollary 2.7 (local property of H-convergence). *If $A_\varepsilon \xrightarrow[\varepsilon\to 0]{H} A$ in a domain Q, then $A_\varepsilon \xrightarrow[\varepsilon\to 0]{H} A$ in any subdomain $Q_1\subset Q$.*

2.2.2. Convergence of energies. In this section, we address a family of Dirichlet problems with nonhomogeneous boundary conditions:

$$\operatorname{div}^\varepsilon_\Lambda\left(\sum_{z'\in\Lambda} a^\varepsilon_{zz'}\,\partial^\varepsilon_{z'}u^\varepsilon\right)=f\,,\ u^\varepsilon-u^0\in W^{1,2}_0(Q_\varepsilon)\,, \tag{2.5}$$

where $u^0\in W^{1,2}(\mathbb{R}^d)$ and $f\in W^{-1,2}(Q)$ are fixed given functions.

We suppose that the family $\{A_\varepsilon\}$ is uniformly elliptic and H-converges to the limit operator A. Then, one can assume without loss of generality that the function u^0 satisfies the equation $A\,u^0=f$ in the domain Q.

In order to show the uniform boundedness of $\{u^\varepsilon\}$ in $W^{1,2}(\overline{Q_\varepsilon})$, we replace u^ε by $u^\varepsilon-u_0$ in (2.5), multiply the resulting equation by $u^\varepsilon-u_0$, and then sum over $\overline{Q}_\varepsilon$. After summation by parts we get

$$\sum_{x\in\overline{Q_\varepsilon}}\sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\partial^\varepsilon_z\big(u^\varepsilon(x)-u_0(x)\big)\,\partial^\varepsilon_{z'}(u^\varepsilon(x)-u_0(x))=\sum_{x\in\overline{Q_\varepsilon}} f(x)\big(u^\varepsilon(x)-u_0(x)\big)$$

$$-\sum_{x\in\overline{Q_\varepsilon}}\sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\partial^\varepsilon_z\big(u^\varepsilon(x)-u_0(x)\big)\,\partial^\varepsilon_{z'}u_0(x).$$

This implies the required boundedness.

By Theorem 2.6 (convergence of arbitrary solution), any weak limiting point of the sequence $\{u^\varepsilon\}$ coincides with u^0 in Q. Hence, the whole family $\{u^\varepsilon\}$ converges to u^0 in $W^{1,2}(\overline{Q_\varepsilon})$ weakly.

Proposition 2.8 (convergence of energies). *Let $A_\varepsilon \xrightarrow[\varepsilon\to 0]{H} A$ and let u^ε be the solution of problem* (2.5). *Then the following limit relation holds true:*

$$\varepsilon^d \sum_{x\in\overline{Q_\varepsilon}} \sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\partial^\varepsilon_z u^\varepsilon(x)\,\partial^\varepsilon_{z'} u^\varepsilon(x) \xrightarrow[\varepsilon\to 0]{} \int_Q \sum_{z,z'\in\Lambda} a_{zz'}(x)\,\frac{\partial}{\partial z}u^0(x)\,\frac{\partial}{\partial z'}u^0(x)\,dx\,.$$

Proof. By (2.5) we have

$$\begin{aligned}
\varepsilon^d &\sum_{x\in\overline{Q_\varepsilon}} \sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\partial^\varepsilon_z (u^\varepsilon-u^0)(x)\,\partial^\varepsilon_{z'}(u^\varepsilon-u^0)(x)\\
&= \varepsilon^d \sum_{x\in\overline{Q_\varepsilon}} \sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\partial^\varepsilon_z u^\varepsilon(x)\,\partial^\varepsilon_{z'}(u^\varepsilon-u^0)(x)\\
&\quad - \varepsilon^d \sum_{x\in\overline{Q_\varepsilon}} \sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\frac{\partial}{\partial z}u^0(x)\,\partial^\varepsilon_{z'}(u^\varepsilon-u^0)(x) + \tau(\varepsilon)\\
&= \varepsilon^d \sum_{x\in Q_\varepsilon} f(x)\,(u^\varepsilon-u^0)(x) - \varepsilon^d \sum_{x\in\overline{Q_\varepsilon}} \sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\frac{\partial}{\partial z}u^0(x)\,\partial^\varepsilon_{z'}u^\varepsilon(x)\\
&\quad + \varepsilon^d \sum_{x\in\overline{Q_\varepsilon}} \sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\frac{\partial}{\partial z}u^0(x)\,\frac{\partial}{\partial z'}u^0(x) \;+\; \tau(\varepsilon)\,;
\end{aligned}$$

here and afterwards $\tau(\varepsilon)$ stands for a generic function that vanishes as $\varepsilon\to 0$. On the other hand,

$$\begin{aligned}
\varepsilon^d &\sum_{x\in\overline{Q_\varepsilon}} \sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\partial^\varepsilon_z (u^\varepsilon-u^0)(x)\,\partial^\varepsilon_{z'}(u^\varepsilon-u^0)(x)\\
&= \varepsilon^d \sum_{x\in\overline{Q_\varepsilon}} \sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\partial^\varepsilon_z u^\varepsilon(x)\,\partial^\varepsilon_{z'}u^\varepsilon(x)\\
&\quad - 2\varepsilon^d \sum_{x\in\overline{Q_\varepsilon}} \sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\frac{\partial}{\partial z}u^0(x)\,\partial^\varepsilon_{z'}u^\varepsilon(x)\\
&\quad + \varepsilon^d \sum_{x\in\overline{Q_\varepsilon}} \sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\frac{\partial}{\partial z}u^0(x)\,\frac{\partial}{\partial z'}u^0(x) + \tau(\varepsilon)\,.
\end{aligned}$$

After subtraction we find

$$\begin{aligned}
\varepsilon^d \sum_{x\in\overline{Q_\varepsilon}} \sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\partial^\varepsilon_z u^\varepsilon(x)\,\partial^\varepsilon_{z'}u^\varepsilon(x) - \varepsilon^d \sum_{x\in\overline{Q_\varepsilon}} \sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\frac{\partial}{\partial z}u^0(x)\,\partial^\varepsilon_{z'}u^\varepsilon(x)\\
- \varepsilon^d \sum_{x\in Q_\varepsilon} f(x)\,(u^\varepsilon-u^0)(x) + \tau(\varepsilon) = 0\,.
\end{aligned} \tag{2.6}$$

Passing to the limit in the last relation, and taking into account the weak convergence of $u^\varepsilon - u^0$ to 0 in $W^{1,2}_0(Q_\varepsilon)$ and the weak convergence of the streams $a^\varepsilon_{zz'}\,\partial^\varepsilon_{z'}u^\varepsilon$ in

$L^2(Q_\varepsilon)$, we obtain

$$\varepsilon^d \sum_{x\in\overline{Q_\varepsilon}} \sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\partial^\varepsilon_z u^\varepsilon(x)\,\partial^\varepsilon_{z'} u^\varepsilon(x) \xrightarrow[\varepsilon\to 0]{} \int_Q \sum_{z,z'\in\Lambda} a_{zz'}(x)\,\frac{\partial}{\partial z} u^0(x)\,\frac{\partial}{\partial z'} u^0(x)\,dx\,. \qquad \square$$

In fact, the result on convergence of energies can be formulated in more "local" form, as follows.

Proposition 2.9. *Under the assumptions of Theorem* 2.6 *one has*

$$\sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\partial_z w^\varepsilon(x)(\partial_{z'} w^\varepsilon(x) + g_{z'}(x)) \xrightarrow[\varepsilon\to 0]{\star} \sum_{z,z'\in\Lambda} a_{zz'}(x)\frac{\partial}{\partial z} w^0(x)\Big(\frac{\partial}{\partial z'} w^0(x) + g_{z'}(x)\Big). \tag{2.7}$$

Proof. In the expression

$$\sum_{z,z'\in\Lambda} \partial_z w^\varepsilon(x) a^\varepsilon_{zz'}(x)(\partial_{z'} w^\varepsilon(x) + g_{z'}(x))\,,$$

the streams $a^\varepsilon_{zz'}(x)(\partial_{z'} w^\varepsilon(x) + g_{z'}(x))$ converge weakly in $L^2(Q_\varepsilon)$ (by Theorem 2.6) to the limit stream $a_{zz'}(x)\Big(\frac{\partial}{\partial z'} w^0(x) + g_{z'}(x)\Big)$, and the family $\partial_z w^\varepsilon(x)$ converges to $\frac{\partial}{\partial z} w^0(x)$ weakly by the assumption of Theorem 2.6. Now, the desired statement follows from Lemma 2.1. $\square$

Remark 2.10. *In the case of elliptic differential equations, H-convergence of operators implies weak L^1-convergence of the corresponding energy functions. This result relies on the Meyers estimates of the gradient of solutions; see Meyers* [20].

For the difference operators the Meyers-type estimates have not been obtained, so the weak L^1-convergence of energies is an open question.

2.2.3. Neumann problem. The notion of the H-limit operator has been expressed in terms of the operators of the corresponding Dirichlet problems. But, as was already mentioned in the previous section, we can also consider other boundary value problems. In this section, the Neumann problem is investigated.

Definition 2.11. *Let $f \in \big(L^2(Q)\big)^{|\Lambda|}$. We say that $u^\varepsilon \in W^{1,2}(Q_\varepsilon)$ is a solution of the Neumann problem for the equation*

$$\mathrm{div}^\varepsilon_\Lambda \left(\sum_{z'\in\Lambda} a^\varepsilon_{zz'}\,\partial^\varepsilon_{z'} u^\varepsilon \right) = \sum_{z\in\Lambda} \partial^\varepsilon_{-z} f^\varepsilon_z$$

if the relation

$$\sum_{x\in\overline{Q_\varepsilon}} \sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\bar\partial^\varepsilon_z \varphi^\varepsilon(x)\,\bar\partial^\varepsilon_{z'} u^\varepsilon(x) = \sum_{x\in\overline{Q_\varepsilon}} \sum_{z\in\Lambda} f^\varepsilon_z(x)\,\bar\partial^\varepsilon_z \varphi^\varepsilon(x) \tag{2.8}$$

holds true for any $\varphi \in W^{1,2}(Q)$; here we use the notation

$$\bar\partial^\varepsilon_z \varphi = \begin{cases} \partial^\varepsilon_z \varphi & \text{if } x + \varepsilon z \in \overline{Q_\varepsilon}, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, the functions u^ε are defined up to an additive constant. To fix the choice of the constant, we assume that $\sum_{x\in\overline{Q_\varepsilon}} u^\varepsilon(x) = 0\,.$

In order to study the Neumann problem, we should modify the definition of uniform ellipticity and impose a slightly stronger condition because Definition 1.1 above does not ensure the coerciveness of problem (2.8).

Definition 2.12. *We say that the family of operators* $\{A_\varepsilon\}$ *is* N-elliptic *in a domain* Q_ε *if the inequality*

$$\sum_{x\in\overline{Q_\varepsilon}}\sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\bar\partial^\varepsilon_z\varphi(x)\bar\partial^\varepsilon_{z'}\varphi(x) \ge c\sum_{x\in\overline{Q_\varepsilon}}\sum_{i=1}^{d}\left(\bar\partial^\varepsilon_{\pm e_i}\varphi(x)\right)^2, \quad c>0, \tag{2.9}$$

holds for any φ.

It should be noted that N-ellipticity implies the uniform ellipticity in the same domain Q and that, under the condition of Proposition 1.3, the family of operators is always N-elliptic.

Example. To clarify the difference between the uniform ellipticity and N-ellipticity we provide below a simple one-dimensional example which shows that due to "boundary effects," a uniformly elliptic operator is not necessary N-elliptic.

Let Q be an open interval $(0,1)$, and suppose $\Lambda=\{0,\pm1,\pm2,\pm3\}$. If we set

$$p_{\pm1}(0)=\frac12,\quad p_z(0)=0 \text{ if } z\ne\pm1;$$

$$p_{-1}(1)=\frac12,\quad p_0(1)=\frac12,\quad p_z(1)=0 \text{ if } z\ne-1,\,0;$$

$$p_{\pm3}(2)=\frac12,\quad p_z(2)=0 \text{ if } z\ne\pm3,$$

and extend this function periodically with period 3, then for $\varepsilon=1/n$ with integer $n>3$ we have

$$Q_\varepsilon=\left\{\frac1n,\frac2n,\dots,\frac{n-1}{n}\right\},\quad \overline{Q}_\varepsilon=\left\{\frac{-2}{n},\frac{-1}{n},0,\frac1n,\frac2n,\dots,\frac{n-1}{n},1,\frac{n+1}{n},\frac{n+2}{n}\right\}.$$

Consider the following test function:

$$\varphi^\varepsilon(x)=\begin{cases}1 & \text{if } x=-\frac2n,-\frac3n,\\ 0 & \text{otherwise.}\end{cases}$$

For this function the LHS of (2.9) is equal to zero while the RHS is strictly positive. Thus (2.9) cannot hold. On the other hand, one can easily verify that this problem is uniformly elliptic.

Proposition 2.13. *Suppose that a family of N-elliptic operators $\{A_\varepsilon\}$ H-converges to the operator A in the domain Q. Then the solutions u^ε of problem (2.8) converge, as $\varepsilon\to0$, in $W^{1,2}(Q_\varepsilon)$ to the solution of the limit Neumann problem: for any $\varphi\in W^{1,2}(Q)$,*

$$\int_Q\left(\sum_{z,z'\in\Lambda}a_{zz'}(x)\frac{\partial}{\partial z}\varphi(x)\,\frac{\partial}{\partial z'}u^0(x)\right)dx=\int_Q\left(\sum_{z\in\Lambda}f_z(x)\frac{\partial}{\partial z}\varphi(x)\right)dx\,.$$

Moreover, the streams also converge.

Proof. Using the Poincaré inequality, we derive from the N-ellipticity the uniform coerciveness of problem (2.8). Thus, the family u^ε is uniformly bounded in $W^{1,2}(Q_\varepsilon)$. By Theorem 2.6, any limit point w^0 of the family u^ε satisfies the H-limit equation and

$$\sum_{z'\in\Lambda} a^\varepsilon_{zz'}\,\partial^\varepsilon_{z'}u^\varepsilon \xrightarrow[\varepsilon\to 0]{} \sum_{z'\in\Lambda} a_{zz'}\,\frac{\partial}{\partial z'}w^0 \quad \text{weakly in } L^2(Q_\varepsilon)\,.$$

So, for any $\varphi \in W^{1,2}(Q)$, passing to the limit in (2.8), we get

$$\int_Q \left(\sum_{z,z'\in\Lambda} a_{zz'}(x)\frac{\partial}{\partial z}\varphi(x)\,\frac{\partial}{\partial z'}w^0(x) \right) dx = \int_Q \left(\sum_{z\in\Lambda} f_z(x)\frac{\partial}{\partial z}\varphi(x) \right) dx\,.$$

Moreover, $\int_Q w^0(x)\,dx = 0$. □

2.2.4. Γ-convergence. The results proved in this section exhibit the relation between the H-convergence of operators and a special kind of convergence of corresponding quadratic forms, so-called Γ-convergence, that was introduced originally in De Giorgi [8].

PROPOSITION 2.14. *Let A_ε be a N-elliptic family of operators in a domain Q. Then, $A_\varepsilon \xrightarrow[\varepsilon\to 0]{H} A$ in Q if and only if the following conditions are satisfied:*

1. *For any $u^0 \in W^{1,2}(Q)$ and for any sequence $w^\varepsilon \in W^{1,2}(Q_\varepsilon)$ such that $w^\varepsilon \xrightarrow[\varepsilon\to 0]{} u^0$ weakly in $W^{1,2}(Q_\varepsilon)$, the following inequality holds:*

$$\begin{aligned}\liminf_{\varepsilon\to 0} \varepsilon^d \sum_{x\in\overline{Q_\varepsilon}} \sum_{z,z'\in\Lambda} & a^\varepsilon_{zz'}(x)\,\bar\partial^\varepsilon_z w^\varepsilon(x)\,\bar\partial^\varepsilon_{z'}w^\varepsilon(x)\\ &\geq \int_Q \sum_{z,z'\in\Lambda} a_{zz'}(x)\,\frac{\partial}{\partial z}u^0(x)\,\frac{\partial}{\partial z'}u^0(x)\,dx\,.\end{aligned}$$

2. *For any $u^0 \in W^{1,2}(Q)$, there exists a sequence $u^\varepsilon \in W^{1,2}(Q_\varepsilon)$ such that $u^\varepsilon \xrightarrow[\varepsilon\to 0]{} u^0$ weakly in $W^{1,2}(Q)$, $u^\varepsilon - u^0 \in W^{1,2}_0(Q)$, and*

$$\begin{aligned}\lim_{\varepsilon\to 0} \varepsilon^d \sum_{x\in\overline{Q_\varepsilon}} \sum_{z,z'\in\Lambda} & a^\varepsilon_{zz'}(x)\,\bar\partial^\varepsilon_z u^\varepsilon(x)\,\bar\partial^\varepsilon_{z'}u^\varepsilon(x)\\ &= \int_Q \sum_{z,z'\in\Lambda} a_{zz'}(x)\,\frac{\partial}{\partial z}u^0(x)\,\frac{\partial}{\partial z'}u^0(x)\,dx\,.\end{aligned}$$

Proof. Suppose that $A_\varepsilon \xrightarrow[\varepsilon\to 0]{H} A\,.$

1. Consider the Neumann problem (2.8), with $f_z = \left(\sum_{z'\in\Lambda} a_{zz'}\,\frac{\partial u^0}{\partial z'}\right)_{z\in\Lambda}$, where u^0 is the solution of the H-limit Neumann problem. The solution u^ε of (2.8) provides the minimum in the following variational problem: $E = \inf_{v\in W^{1,2}(Q_\varepsilon)} J^\varepsilon(v)$, where

$$J^\varepsilon(v) = \varepsilon^d \sum_{x\in\overline{Q_\varepsilon}} \sum_{z,z'\in\Lambda} \left[a^\varepsilon_{zz'}(x)\,\bar\partial^\varepsilon_z v(x)\,\bar\partial^\varepsilon_{z'}v(x) - 2a_{zz'}(x)\,\bar\partial^\varepsilon_z v(x)\,\frac{\partial}{\partial z'}u^0(x)\right]\,.$$

For any sequence $\{w^\varepsilon\}$ such that $w^\varepsilon \to u^0$ weakly in $W^{1,2}(Q_\varepsilon)$, we have

$$J^\varepsilon(w^\varepsilon) \geq J^\varepsilon(u^\varepsilon)\,. \tag{2.10}$$

Then, by Proposition (2.13), $\partial^\varepsilon_z u^\varepsilon \xrightarrow[\varepsilon\to 0]{} \dfrac{\partial}{\partial z}u^0$ weakly in $L^2(Q_\varepsilon)$ and, therefore,

$$\begin{aligned}
\varepsilon^d \sum_{x\in\overline{Q_\varepsilon}}\sum_{z,z'\in\Lambda} & a^\varepsilon_{zz'}(x)\,\bar\partial^\varepsilon_z w^\varepsilon(x)\,\bar\partial^\varepsilon_{z'} w^\varepsilon(x) \\
&= J^\varepsilon(w^\varepsilon) + 2\sum_{x\in\overline{Q_\varepsilon}}\sum_{z,z'\in\Lambda} a_{zz'}(x)\,\bar\partial^\varepsilon_z w^\varepsilon(x)\,\frac{\partial}{\partial z'}u^0(x) \\
&\geq J^\varepsilon(u^\varepsilon) + 2\sum_{x\in\overline{Q_\varepsilon}}\sum_{z,z'\in\Lambda} a_{zz'}(x)\,\bar\partial^\varepsilon_z w^\varepsilon(x)\,\frac{\partial}{\partial z'}u^0(x) \\
&= -\sum_{x\in\overline{Q_\varepsilon}}\sum_{z,z'\in\Lambda} a_{zz'}(x)\,\bar\partial^\varepsilon_z u^\varepsilon(x)\,\frac{\partial}{\partial z'}u^0(x) \\
&\quad +2\sum_{x\in\overline{Q_\varepsilon}}\sum_{z,z'\in\Lambda} a_{zz'}(x)\,\bar\partial^\varepsilon_z w^\varepsilon(x)\,\frac{\partial}{\partial z'}u^0(x) \\
&\xrightarrow[\varepsilon\to 0]{} \int_Q \sum_{z,z'\in\Lambda} a_{zz'}(x)\,\frac{\partial}{\partial z}u^0(x)\,\frac{\partial}{\partial z'}u^0(x)\,dx\,.
\end{aligned}$$

Equation (2.8) has also been used here. Now, taking the infimum limit in both sides of (2.10), we obtain the required inequality.

2. It is the statement of Proposition 2.8.

The remaining part of the proposition is an easy consequence of the uniqueness of the H-limit. □

REMARK 2.15. *The statements of Propositions* 2.8 *and* 2.14 *remain valid if we replace the sums over* $x \in \overline{Q_\varepsilon}$ *by the sums over* $x \in Q_\varepsilon$.

2.3. Description of the random environment.

In this section we introduce random difference elliptic operators with statistically homogeneous rapidly oscillating coefficients.

Let $(\Omega, \mathcal{F}, \mu)$ be a standard probability space, where $\mathcal{F}$ is a σ-algebra of subsets of Ω and μ is a probability measure. Let $\{T_x : \Omega \mapsto \Omega;\, x \in \mathbb{Z}^d\}$ be a group of $\mathcal{F}$-measurable transformations which preserve the measure μ:

1. $T_x : \Omega \mapsto \Omega$ is $\mathcal{F}$–measurable for all $x \in \mathbb{Z}^d$,
2. $\mu(T_x\mathcal{B}) = \mu(\mathcal{B})$, for any $\mathcal{B} \in \mathcal{F}$ and $x \in \mathbb{Z}^d$,
3. $T_0 = I\,,\ T_x \circ T_y = T_{x+y}$.

In what follows we assume that the group T_x is ergodic. That is, any $f \in L^1(\Omega)$ such that $f(T_x\,\omega) = f(\omega)$ μ-a.s for each $x \in \mathbb{Z}^d$ is equal to a constant μ-a.s.

Let Λ be a finite subset of $\mathbb{Z}^d$. Given a matrix-valued $\mathcal{F}$-measurable function $\{a_{zz'}(\omega)\}$, $z, z' \in \Lambda$, with values in the space of symmetric $|\Lambda| \times |\Lambda|$ matrices, we define a family of difference operators A^ε with the coefficients

$$a^\varepsilon_{zz'}(x) = a_{zz'}(T_{x/\varepsilon}\omega), \quad x \in \varepsilon\mathbb{Z}^d,\ z, z' \in \Lambda\,. \tag{2.11}$$

We suppose here that $\pm e_i \in \Lambda,\ i = 1, \ldots, d$, and that

$$\sum_{z,z' \in \Lambda} a_{zz'}(\omega)\eta_z\eta_{z'} \geq c\sum_{i=1}^{d} |\eta_{\pm e_i}|^2\,, \quad \eta \in \mathbb{R}^{|\Lambda|}\,. \tag{2.12}$$

$$|a_{zz'}(\omega)| \leq c_1, \quad z, z' \in \Lambda. \tag{2.13}$$

It is easy to see that these inequalities imply the N-ellipticity and the uniform ellipticity of the corresponding family A_ε in any regular domain Q.

In applications, especially in those related to random walks, we usually deal with the following particular case of the above construction.

Let $\{q(\omega, z),\ z \in \mathbb{Z}^d\}$ be a family of random variables such that μ-a.s,

1. $\sum\limits_{z \in \mathbb{Z}^d} q(\omega, z) = 1\,,$
2. $q(T_x\omega, z) = q(T_{x+z}\omega, -z)\,,$
3. $q(\omega, z) \geq 0, \quad q(\omega, \pm e_i) \geq \delta > 0,\ i = 1, \ldots, d$ (ellipticity condition).

We introduce a family of transition probabilities as follows:

$$p_z(x) = q(T_x\omega, z),$$

where the argument ω, treated as a realization of the medium, is omitted. The important characteristic of a family of transition probabilities is the structure of its support:

$$\Lambda = \left\{ z \in \mathbb{Z}^d \mid \operatorname*{ess\,sup}_{\Omega} p_z(x) \neq 0 \right\}.$$

In all the models considered below, the set Λ is finite.

Now, if we denote $p_z^\varepsilon(x) = p_z(\varepsilon^{-1}x),\ x \in Q_\varepsilon,\ z \in \Lambda$, then due to the assumptions on $q(\omega, x)$, problem (1.5) is uniformly and N-elliptic.

It is convenient to define the "ω-divergence" operator:

for any random variable $v \in L^2(\Omega)\,, \quad \operatorname{div}_\omega v(\omega) \overset{\triangle}{=} \sum\limits_{z \in \Lambda} v(T_{-z}\omega) - v(\omega)\,.$

We will use it in the following analysis.

2.4. Homogenization of random operators. This section is devoted to homogenization of the random difference operators introduced in the preceding section. The first proof of the homogenization theorem for such operators was obtained in [15], where the "corrector technique" was used. Here we give another proof of the theorem, which relies on the compensated compactness lemma.

2.4.1. Auxiliary problem. Let us define the following subspaces of $\left(L^2(\Omega)\right)^{|\Lambda|}$ (see Kozlov [15]):

$L^2_{pot}(\Omega, \Lambda)$ is the closure of the set
$\left\{v \in (L^2(\Omega))^{|\Lambda|}\,;\, v_z(\omega) = u(T_z\omega) - u(\omega) \text{ for some } u \in L^\infty(\Omega)\right\}$,
$L^2_{sol}(\Omega, \Lambda)$ is the closure of the set: $\left\{v \in (L^2(\Omega))^{|\Lambda|}\,;\, \operatorname{div}_\omega\, v = 0\right\}$.

For $\lambda \in \mathbb{R}^{|\Lambda|}$ we denote by $\mathcal{V}^2_{pot,\lambda}(\Omega, \Lambda)$ the closed set $\left\{v + \lambda\,;\, v \in L^2_{pot}(\Omega, \Lambda)\right\}$.

Consider the following auxiliary problem: given $\lambda \in \mathbb{R}^{|\Lambda|}$, find $v \in \mathcal{V}^2_{pot,\lambda}(\Omega, \Lambda)$ such that

$$\mathrm{div}_\omega \left(\sum_{z' \in \Lambda} a_{zz'}(\omega)\, v_{z'}(\omega) \right) = 0\,. \tag{2.14}$$

In order to prove the existence and uniqueness of the solution of this problem we introduce the operator

$$\begin{array}{rcl} A_{pot}\ :\ \ L^2_{pot}(\Omega,\Lambda) & \mapsto & L^2_{pot}(\Omega,\Lambda) \\ (v_z)_{z\in\Lambda} & \mapsto & \Pi_{pot}\left(\displaystyle\sum_{z'\in\Lambda} a_{zz'}(\omega)\, v_{z'}(\omega) \right), \end{array}$$

where Π_{pot} is the orthogonal projection onto the subspace $L^2_{pot}(\Omega, \Lambda)$.

In view of the Weyl decomposition (see Kozlov [15]) $\left(L^2(\Omega)\right)^{|\Lambda|} = L^2_{pot}(\Omega, \Lambda) \oplus L^2_{sol}(\Omega, \Lambda)$, we can rewrite the problem (2.14) in the following form: given $\lambda \in \mathbb{R}^{|\Lambda|}$, find $v \in L^2_{pot}(\Omega)$ such that

$$A_{\rm pot} v = \Pi_{pot} \left(\sum_{z' \in \Lambda} a_{zz'}(\omega)\, \lambda_{z'} \right).$$

The operator A_{pot} is coercive. Indeed, for any $v \in L^2_{pot}(\Omega, \Lambda)$, we have

$$\begin{aligned} (A_{pot} v, v) &= \sum_{z\in\Lambda} \left(\Pi_{pot} \left(\sum_{z'\in\Lambda} a_{zz'}(\omega)\, v_{z'}(\omega) \right),\ v_z(\omega) \right)_{L^2(\Omega)} \\ &= \sum_{z\in\Lambda} \left(\sum_{z'\in\Lambda} a_{zz'}(\omega)\, v_{z'}(\omega),\ \Pi_{pot}\left(v_z(\omega)\right) \right)_{L^2(\Omega)} \\ &= \sum_{z\in\Lambda} \left(\sum_{z'\in\Lambda} a_{zz'}(\omega)\, v_{z'}(\omega),\ v_z(\omega) \right)_{L^2(\Omega)}. \end{aligned}$$

According to hypothesis (2.12), this implies $(A_{\rm pot} v, v) \ge c\, E \left[\sum_{i=1}^n |v_{\pm e_i}|^2\right]$, where E stands for the expectation with respect to the measure μ. On the other hand, for any v of the form $v_z(\omega) = u(T_z\omega) - u(\omega)$, $u \in L^2(\Omega)$, we have

$$\begin{aligned} \|v\|^2_{(L^2(\Omega))^{|\Lambda|}} &= E\left[\sum_{z\in\Lambda} |v_z(\omega)|^2 \right] \ =\ E\left[\sum_{z\in\Lambda} |u(T_z\omega) - u(\omega)|^2 \right] \\ &= E\left[\sum_{z\in\Lambda} \left| \sum_{i=0}^{N(z)-1} u(T_{\zeta_{i+1}}\omega) - u(T_{\zeta_i}\omega) \right|^2 \right], \end{aligned}$$

where $\zeta_0 = 0$, $\zeta_{N(z)} = z$, $|\zeta_{i+1} - \zeta_i| = 1$, and $N(z) \le d\,\mathrm{diam}(\Lambda)$. Therefore,

$$\|v\|^2_{(L^2(\Omega))^{|\Lambda|}} \le E\left[d(\mathrm{diam}(\Lambda))^2 |\Lambda| \sum_{i=1}^d (v_{\pm e_i}(\omega))^2 \right] \le c_1(d,\Lambda)\, E\left[\sum_{i=1}^d (v_{\pm e_i}(\omega))^2 \right].$$

By definition, the said set of $v(\omega)$ is dense in $L^2_{pot}(\Omega, \Lambda)$, and by the continuity arguments, the latter estimate holds for any $v \in L^2_{pot}(\Omega, \Lambda)$.

Thus, $(A_{\rm pot} v, v) \ge c_2(d, \Lambda)\, \|v\|^2_{(L^2(\Omega))^{|\Lambda|}}$, and the desired existence and uniqueness follow from the Lax–Milgram lemma.

2.4.2. Homogenization. In this section, we study the family of random operators $\{A_\varepsilon\}$ with statistically homogeneous coefficients given by (2.11). The homogenization theorem for such operators was originally proved in [15]. We give another proof based on the compensated compactness lemma, which seems to be easier and shorter. The main result here is the following theorem.

THEOREM 2.16. *Let the coefficients of A_ε be given by* (2.11), *and suppose the condition* (2.12) *is fulfilled. Then, a.s., the family $\{A_\varepsilon\}$ admits homogenization and the limit matrix $\mathcal{A}^0$ does not depend on ω.*

Proof. For a fixed $f \in W^{-1,2}(Q)$, consider the following Dirichlet problems:

$$\operatorname{div}_\Lambda^\varepsilon \left(\sum_{z' \in \Lambda} a^\varepsilon_{zz'} \partial^\varepsilon_{z'} u^\varepsilon \right) = f\,, \quad u^\varepsilon \in W_0^{1,2}(Q_\varepsilon)\,. \tag{2.15}$$

Since u^ε and $\sum_{z'\in\Lambda} a^\varepsilon_{zz'}\, \partial^\varepsilon_{z'} u^\varepsilon$ are uniformly bounded, respectively, in $W^{1,2}(Q_\varepsilon)$ and $(L^2(Q_\varepsilon))^{|\Lambda|}$, we have

$$u^\varepsilon \xrightarrow[\varepsilon\to 0]{} u^0 \quad \text{weakly in } W_0^{1,2}(Q_\varepsilon)\,,$$
$$s^\varepsilon \xrightarrow[\varepsilon\to 0]{} s^0 \quad \text{weakly in } \left(L^2(Q_\varepsilon)\right)^{|\Lambda|}\,,$$

where s^ε_z stands for $\sum_{z'\in\Lambda} a^\varepsilon_{zz'}\, \partial^\varepsilon_{z'} u^\varepsilon$.

Let $v_z(\omega)$ solve the auxiliary problem (2.14). If we denote $v^\varepsilon(x) \triangleq v\left(T_{x/\varepsilon}\,\omega\right)$, $q^\varepsilon \triangleq v^\varepsilon \mathcal{A}^\varepsilon$, i.e., $\forall z \in \Lambda$, $q^\varepsilon_z = \sum_{z'\in\Lambda} v^\varepsilon_{z'}\, a^\varepsilon_{zz'}$, then, the identity

$$\sum_{x\in Q_\varepsilon} \sum_{z\in\Lambda} s^\varepsilon_z(x)\, v^\varepsilon_z(x) = \sum_{x\in Q_\varepsilon} \sum_{z\in\Lambda} q^\varepsilon_z(x)\, \partial^\varepsilon_z u^\varepsilon(x) \tag{2.16}$$

obviously holds. We introduce a constant matrix $\mathcal{A}$ to satisfy the relation $E(q^\varepsilon) = \lambda\, \mathcal{A}^0$. This matrix is well defined because q^ε is a linear functional of λ. By the Birkhoff ergodic theorem, we have

$$v^\varepsilon \xrightarrow[\varepsilon\to 0]{} E(v^\varepsilon) = \lambda \quad \text{weakly in } L^2(Q_\varepsilon) \text{ a.s.},$$
$$q^\varepsilon \xrightarrow[\varepsilon\to 0]{} E(q^\varepsilon) = \lambda\, \mathcal{A} \quad \text{weakly in } L^2(Q_\varepsilon) \text{ a.s.}$$

It follows from (2.14) and the definition of $\operatorname{div}_\omega$ and $\operatorname{div}^\varepsilon_\Lambda$ that for almost all realizations we have $\operatorname{div}^\varepsilon_\Lambda q^\varepsilon = 0$, while the fact that $v - \lambda \in L^2_{pot}(\Omega, \Lambda)$ implies a.s. the relation $v^\varepsilon_z = \partial^\varepsilon_z \theta^\varepsilon$ for some (in general not statistically homogeneous) functions θ^ε. Also, from (2.15) we have $\operatorname{div}^\varepsilon_\Lambda s^\varepsilon = f$. By Lemma 2.1,

$$\sum_{z\in\Lambda} s^\varepsilon_z\, v^\varepsilon_z \xrightarrow[\varepsilon\to 0]{\star} \sum_{z\in\Lambda} s^0_z\, \lambda_z$$

and

$$\sum_{z\in\Lambda} q^\varepsilon_z\, \partial^\varepsilon_z u^\varepsilon \xrightarrow[\varepsilon\to 0]{\star} \sum_{z,z'\in\Lambda} \lambda_z\, a^0_{zz'}\, \frac{\partial}{\partial z'} u^0\,,$$

or, equivalently,

$$\sum_{z,z'\in\Lambda} v^\varepsilon_{z'}\, a^\varepsilon_{zz'}\, \partial^\varepsilon_z u^\varepsilon \xrightarrow[\varepsilon\to 0]{\star} \sum_{z,z'\in\Lambda} \lambda_z\, a^0_{zz'}\, \frac{\partial}{\partial z'} u^0;$$

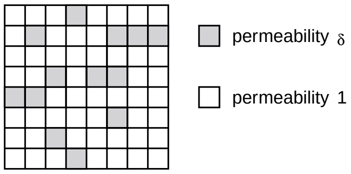


FIG. 3.1. *Example of a realization of the random medium.*

we have also used here Proposition 3 from [15]. Hence, passing to the limit in (2.16) and bearing in mind the fact that λ is an arbitrary vector, we find

$$s_z^0 = \sum_{z' \in \Lambda} a^0_{zz'} \, \frac{\partial}{\partial z'} u^0 \, .$$

Since $\sum_{z\in\Lambda} \frac{\partial}{\partial z} s_z^0 = f$, the function u^0 is the solution of the homogenized problem and $\mathcal{A}^0$ is the limit matrix. □

3. Asymptotic behavior of the effective coefficient. In this second part of the work, we consider the difference operators obtained by discretizing a random two-dimensional high-contrast checker-board structure, as various discretization procedures are applied. For each discretization method, we find the asymptotics of the effective coefficient. The results obtained in this section rely essentially on the fine results from the percolation theory, such as channel property and related statements. For the reader's convenience, we formulate these results and provide necessary definitions in section 3.1.

To define the random media, we split the plane $\mathbb{R}^2$ into regular squares $\{[-\frac{1}{2}, \frac{1}{2}]^2 + j\}$, $j \in \mathbb{Z}^2$, and assign a value of permeability, independently at each square, as follows:

$$\kappa(y) \triangleq \left\{ \begin{array}{lll} \delta & \text{with probability} & p \\ 1 & \text{with probability} & 1-p \end{array} \right. , \quad y \in \left[-\frac{1}{2}, \frac{1}{2} \right]^2 + j, \; j \in \mathbb{Z}^2,$$

where δ is a small strictly positive parameter (see Figure 3.1). Then, we consider the grid $\mathbb{Z}^2$, fix a finite set $\Lambda \subset \mathbb{Z}^2$, and define the transition probabilities $\{p_z(x);\, x \in \mathbb{Z}^2,\, z \in \Lambda\}$ to be a function of $\{\kappa(x+z)\}$, $z \in \Lambda$. Finally, we define the coefficients of operator A_ε in terms of $\{p_z(x)\}$ by (1.6).

Henceforth, we suppose that the properties (1), (2), and (3) in section 1.1 are satisfied. It then follows from the independence of $\kappa(j)$ for different $j \in \mathbb{Z}^2$ that the family $\{p_z(x)\}$ is ergodic. Now, the following assertion is a direct consequence of Theorem 2.16 (see also Kozlov [15, section 2]).

PROPOSITION 3.1.

1. *The operators A_ε G-converge as $\varepsilon \to 0$ to an elliptic operator with constant nonrandom coefficients $\mathcal{A} = \{a_{zz'}\}_{z,z' \in \Lambda \setminus (0,0)}$.*
2. *The limit matrix is isotropic: $\mathcal{A} = a^\delta(p)\, I$ (I is the identity matrix).*

We call $a^\delta(p)$ the effective coefficient and study its asymptotics as $\delta \to 0$ for various $p \in [0, 1]$.

3.1. Some results from percolation theory. In this section, we quote and discuss briefly several results from percolation theory. We consider the so-called site percolation model (see Grimmett [10]) and, following the tradition, say black and

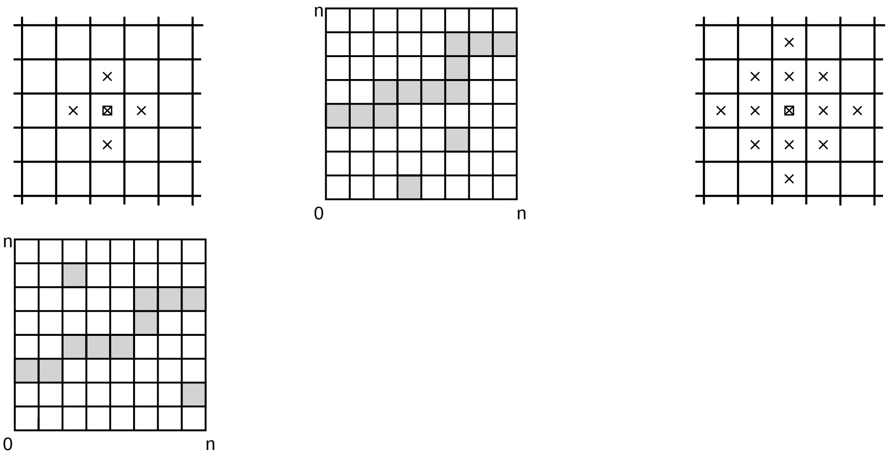


Fig. 3.2. *The neighbor squares and black channels in the cases $\gamma = 1$ (left) and $\gamma = \sqrt{2}$ (right).*

white squares instead of "δ" and "1" squares, respectively. All the squares are enumerated by the coordinates of their centers and the distance $\operatorname{dist}(i, j)$ between squares i and j, $(i, j \in \mathbb{Z}^2)$, is defined as the Euclidean distance $|i - j|$.

Definition 3.2.

- *Two black squares i and j are* γ-connected *if* $\operatorname{dist}(i, j) \leq \gamma$. *As soon as the value of γ is fixed, we just refer to* connected *squares or* neighbor *squares.*
- *Consider the random subgraph containing only the black squares. The connected components of this graph are called* black clusters.
- *A finite set of black squares forms a* black channel *if the squares can be enumerated in such a way that any two successive squares in this enumeration are γ-connected (see Figure* 3.2 *for examples).*

Similarly, we define γ-connected white squares, white clusters and white channels. When the probability p varies, the geometric properties of the black clusters are modified. The more p increases, the bigger are the sizes of the clusters, and they eventually form the unique infinite cluster (see, for example, Grimmett [10]). Below, some basic constructions of percolation theory are presented.

The probability space is introduced as follows. As sample space, we take $K = \Pi_{s \in \mathbb{Z}^2}\{\delta, 1\}$. Each point of K: $\kappa = (\kappa(s); s \in \mathbb{Z}^2)$ is called a configuration. We take $\mathcal{G}$ to be the σ-field of subsets of K generated by the finite dimensional cylinders. And, for each $p \in [0, 1]$, we define the probability measure P_p as the product measure on $(K, \mathcal{G})$ such that the random variables $\kappa(s), s \in \mathbb{Z}^2$ are independent and satisfy $P_p(\kappa(x) = \delta) = p$.

In what follows we identify the probability space $(K, \mathcal{G}, P_p)$ with the general probability space $(\Omega, \mathcal{F}, P)$ defined above.

Let $|C|$ be the cardinal of the cluster which contains the origin. The cluster-size distribution is given by $\theta_n(p) = P_p(|C| = n), \ n \in \mathbb{N}^*$. The probability $\theta(p) = P_p(|C| = +\infty)$ that the origin belongs to the infinite cluster is called the *percolation probability.* There exists a *critical probability* $p_c(\gamma)$, also called the *percolation threshold*, such that

$$\begin{cases} \theta(p) = 0 & \text{if} \quad p \leq p_c(\gamma)\,, \\ \theta(p) > 0 & \text{if} \quad p > p_c(\gamma)\,. \end{cases}$$

Table 3.1
Evolution of the number of infinite cluster with respect to p.

<table>
<tr><th>p</th><th>0</th><th>$p_c(2)$</th><th>$p_c(\sqrt{2})$</th><th>$p_c(1)$</th><th>$1-p_c(2)$</th><th>1</th></tr>
<tr><td>$\gamma = 1$</td><td colspan="2">White</td><td>No infinite cluster</td><td colspan="3">Black</td></tr>
<tr><td>$\gamma = \sqrt{2}$</td><td colspan="2">White</td><td>Black and White</td><td colspan="3">Black</td></tr>
<tr><td>$\gamma = 2$</td><td>White</td><td colspan="3">Black and White</td><td colspan="2">Black</td></tr>
</table>

Thus, for each fixed γ, the *critical probability* is $p_c(\gamma) \overset{\triangle}{=} \sup\{p : \theta(p) = 0\}$.

Figure 3.2 shows the sets of neighbor squares, with respect to the marked square, in the cases $\gamma = 1$ and $\gamma = \sqrt{2}$, and it emphasizes the difference between the structures of channels.

In Table 3.1, we can see, for three different values of γ, the presence of white and black clusters with respect to the values of p. The following relation holds: $p_c(1) + p_c(\sqrt{2}) = 1$, while $p_c(1) \sim 0.59$ and $p_c(\sqrt{2}) \sim 0.41$ (see Kesten [12]).

Moreover, according to Aizenman and Grimmett [1], $p_c(2) < p_c(\sqrt{2})$.

3.1.1. The channel property. Denote by $N(n)$ the number of mutually non-intersecting black channels joining the left and the right sides of the box $[0, n]^2$.

Proposition 3.3 (see Kesten [12, section 11]). *Let $\gamma = 1$ or $\gamma = \sqrt{2}$. If $p > p_c(\gamma)$, then for almost all $\kappa \in K$ the inequality*

$$N(n) \geq c(p)\, n\,, \quad c(p) > 0,$$

holds for any $n \geq n_0(\kappa)$

Remark 3.4. *In fact, this result holds true for any value of γ (see Golden and Kozlov* [9]*).*

Remark 3.5. *For all $\gamma \geq \sqrt{2}$, the percolation models admit the coexistence of the channels of both colors (see Figure* 3.2*).* The geometry of the white and black subgraphs is rather different in subcritical and supercritical zones. In this connection, it is interesting to study carefully what happens near $p_c(\gamma)$.

Proposition 3.6 (see Kesten [12, section 11]). *There exist some strictly positive constants c_1, c_2, c_3, δ_1, δ_2 such that, for $p > p_c(\gamma)$,*

$$P_p\left(N(n) \geq c_1\,(p - p_c(\gamma))^{\delta_1} n\right) \geq 1 - c_2\,(n+1)\,e^{-c_3\, n\,(p-p_c(\gamma))^{\alpha_2}}\,.$$

By the Borel–Cantelli lemma, we have

$$c(p) \geq c_1\,(p - p_c(\gamma))^{\delta_1}\,. \tag{3.1}$$

Remark 3.7. *One can easily check that all the channels can be chosen to be no longer than $\theta(p)\, n$.*

3.2. Behavior of the effective coefficient. In this section, for the checkerboard model introduced above, we consider several discrete models characterized by

- the set of admissible jumps, i.e., the set Λ;
- the corresponding transition probabilities $\{p_z\}_{z\in\Lambda}$.

In all these models, the distribution of $\{p_z\}_{z\in\Lambda}$ will be invariant with respect to rotations at the angle $\pi/2$. This symmetry implies the isotropy of the effective tensor, and thus there is only one scalar effective coefficient $a^\delta(p)$ to be determined.

For each model, we study the limit behavior of the effective coefficient as $\delta \to 0$.

3.2.1. Harmonic mean. We begin by considering the "harmonic mean" model. Namely, we assume that

$$\Lambda = \{\pm(1,0),\, \pm(0,1),\, (0,0)\}$$

and define the transition probabilities as the harmonic mean of the values of $\kappa(\cdot)$ at the corresponding points:

$$p_z(x) = \begin{cases} \dfrac{1}{4}\,\dfrac{2\,\kappa(x)\,\kappa(x+z)}{(\kappa(x)+\kappa(x+z))} & \text{if } z \in \Lambda \setminus \{(0,0)\}\,, \\ 1 - \sum_{z\in\Lambda\setminus\{(0,0)\}} p_z(x) & \text{if } z = (0,0)\,, \\ 0 & \text{if } z \notin \Lambda\,. \end{cases}$$

Clearly, the family $\{p_z(x)\}$ satisfies the conditions (1), (2), and (3) in section 1.1, and moreover, its distribution is isotropic.

Remark 3.8. *The choice of the harmonic mean is natural in the framework of the finite volume approach. Indeed, with this choice for the coefficients, we* conserve the fluxes. *This conservation is violated under another choices (see explanations in McCarthy* [19]*).*

The asymptotic behavior of the effective coefficient $a^\delta(p)$ as $\delta \to 0$ is described by the following statement.

Theorem 3.9. *The effective coefficient $a^\delta(p)$ satisfies, for small δ, the following inequalities:*

$$\begin{array}{ll} 0 < c_1(p) \le a^\delta(p) \le 1 & \text{if } \ 0 \le p < p_c(\sqrt{2}), \\ \delta \le a^\delta(p) \le c_2(p)\,\delta,\ c_2(p) > 0 & \text{if } \ p_c(\sqrt{2}) < p \le 1. \end{array}$$

This means, in particular, that $a^\delta(p)$ does not vanish as $\delta \to 0$ if $p < p_c(\sqrt{2})$.

Proof.

1. *Case* $0 \le p < p_c(\sqrt{2})$.

Consider the percolation model with $\gamma = 1$. By Proposition 3.3, for $0 \le p < 1 - p_c(1)$ there are at least $N(n) = c(p)\,n$ mutually nonintersecting white channels joining the left and the right sides of the square $[0,n]^2$. We denote by C_k the kth channel, $1 \le k \le N(n)$.

Define on the space $\big(L^2(\Omega)\big)^{|\Lambda|}$ the following seminorm:

$$\|\phi\|^2 \overset{\triangle}{=} E\left\{\sum_{z\in\Lambda} p_z(\omega)\,(\phi_z(\omega))^2\right\}, \tag{3.2}$$

where E is the expectation related to the measure μ. In fact, under the assumptions of the theorem, it is a norm, but we will not use this fact.

Let $P_1(z) = z_1$ be the projection onto the first coordinate of vector z. According to Kozlov ([14, Chapter II, section 2], the effective coefficient $a^\delta(p)$ can be calculated as follows:

$$a^\delta(p) = \inf_{\varphi\in L^2_{pot}(\Omega,\Lambda)} \|P_1(z) - \varphi\|^2\,, \tag{3.3}$$

where the subspace $L^2_{pot}(\Omega,\Lambda)$ has been defined in section 2.4.1 of this paper. Denote by $\mathcal{H}$ the linear set $\mathcal{H} = \{\varphi_z(\omega) = \tilde\varphi(T_z\omega) - \tilde\varphi(\omega)\,;\ \tilde\varphi \in L^\infty(\Omega)\}$. This set $\mathcal{H}$ is dense in $L^2_{pot}(\Omega,\Lambda)$ (see section 2.4.1) and the functional $\varphi \to \|z_1 - \varphi\|$ is continuous in $L^2_{pot}(\Omega,\Lambda)$. Therefore, the infimum over $L^2_{pot}(\Omega,\Lambda)$ in (3.3) can be replaced by the infimum over $\mathcal{H}$.

Let φ belong to $\mathcal{H}$: there exists $\tilde{\varphi} \in L^\infty(\Omega)$ such that $\varphi_z(\omega) = \tilde{\varphi}(T_z\,\omega) - \tilde{\varphi}(\omega)$. Then,

$$\|P_1(z) - \varphi\|^2 = E\left\{\sum_{z\in\Lambda} p_z(\omega)(z_1 - (\tilde{\phi}(T_z\,\omega) - \tilde{\phi}(\omega)))^2\right\}.$$

Since T_x is ergodic, by the Birkhoff theorem we have for almost all realizations

$$\begin{aligned}\|P_1(z) - \phi\|^2 &= \lim_{n\to+\infty} \frac{1}{n^2} \sum_{x\in\mathbb{Z}^2\cap[0,n]^2} \sum_{z\in\Lambda} p_z(T_x\,\omega)\,\left(z_1 - \tilde{\phi}(T_{x+z}\,\omega) + \tilde{\phi}(T_x\,\omega)\right)^2 \\ &= \lim_{n\to+\infty} \frac{1}{n^2} \sum_{x\in\mathbb{Z}^2\cap[0,n]^2} \sum_{z\in\Lambda} p_z(x)\,\left(z_1 - \tilde{\phi}(T_{x+z}\,\omega) + \tilde{\phi}(T_x\,\omega)\right)^2 . \qquad (3.4)\end{aligned}$$

Our goal now is to construct a uniformly positive lower bound for $a^\delta(p)$. To this end, on the RHS of the last formula, we first take into account only the points x located inside the channels:

$$\|P_1(z) - \phi\|^2 \geq \liminf_{n\to+\infty} \frac{1}{n^2} \sum_{x\in C}\sum_{z\in\Lambda} p_z(x)\,\left(z_1 - \tilde{\phi}(T_{x+z}\,\omega) + \tilde{\phi}(T_x\,\omega)\right)^2 ,$$

where C stands for the union of white channels. Then, we enumerate the points x along each channel in such a way that any consecutive numbers correspond to neighbor points, and we replace the inner sum over $z \in \Lambda(x)$ by the sum over z such that $x+z$ belong to the same channel as x and have greater index than x. Denote this latter set of z by $\lambda(x)$, and notice that for each x from the union of white channels $\lambda(x)$ is not empty and consists of only one element. For $z \in \lambda(x)$, we clearly have $p_z(x) = 1/4$. Hence,

$$\|P_1 - \varphi\|^2 \geq \liminf_{n\to+\infty} \frac{1}{4\,n^2} \sum_{x\in C}\ \sum_{z\in\lambda(x)} \left(z_1 - \tilde{\phi}(T_{x+z}\,\omega) + \tilde{\phi}(T_x\,\omega)\right)^2 ,$$

If we denote $S(x) = \sum_{z\in\lambda(x)} (z_1 - \tilde{e}(T_{x+z}\,\omega) + \tilde{e}(T_x\,\omega))$, and enumerate the channels $C = \cup_{k=1}^{N(n)} C_k$, then, for the kth channel, we have

$$\begin{aligned}\sum_{x\in C_k} S(x) &= \sum_{x\in C_k}\sum_{z\in\lambda(x)} (z_1 - \tilde{\phi}(T_{x+z}\omega) + \tilde{\phi}(T_x\omega)) \\ &= n + \sum_{x\in C_k}\sum_{z\in\lambda(x)} (-\tilde{\phi}(T_{x+z}\omega) + \tilde{\phi}(T_x\omega)) \\ &= n + \tilde{\phi}(T_{x_s(C_k)}\omega) - \tilde{\phi}(T_{x_f(C_k)}\,\omega) \geq n - c\,,\end{aligned}$$

where $c = 2\,\|\tilde{\phi}\|_{L^\infty(\Omega)}$, and $x_s(C_k)$ and $x_f(C_k)$ are, respectively, the starting and final points of kth channel. Summing up over the channels, we obtain

$$\sum_{x\in C} S(x) \geq (n-c)\,N(n)\,. \qquad (3.5)$$

By the Cauchy inequality, taking into account Remark 3.7, we get

$$\sum_{x\in C} S(x)^2 \geq \frac{\left(\sum_{x\in C} S(x)\right)^2}{\theta(p)\,n\,N(n)}\,.$$

In view of (3.5) this implies

$$\sum_{x\in C} S(x)^2 \ge \frac{(n-c)^2\, N(n)^2}{\theta(p)\, n\, N(n)}\,, \tag{3.6}$$

and

$$\|P_1(z)-\phi\|^2 \ge \lim_{n\to+\infty} \frac{1}{4\,n^2}\frac{c(p)}{\theta(p)}\,(n-c)^2 \ge c_1 > 0\,. \tag{3.7}$$

Hence: $a^\delta(p) \ge c_1 > 0$. The upper bound $a^\delta(p) \le 1$ is obvious and, finally,

$$0 < c_1 \le a^\delta(p) \le 1\,.$$

2. *Case* $p_c(\sqrt{2}) < p \le 1$.

Consider the percolation model with $\gamma = \sqrt{2}$. There are at least $c(p)\,n$ non-intersecting black channels $C_k,\ k = 1, 2, \dots, N(n)$, joining the left and the right sides of the square $[0, n]^2$ (see Proposition 3.3).

Let us denote $\varepsilon = 1/n$ and define functions w^ε on $\varepsilon\mathbb{Z}^2 \cap [0, 1]^2$ as follows:

- $w^\varepsilon(\cdot, 0) = 0\,,\ w^\varepsilon(\cdot, 1) = 1$ (boundary conditions),
- $w^\varepsilon(x) = \frac{(k-1/2)}{N(n)}$ for $x \in \varepsilon C_k$,
- $w^\varepsilon(x) = \frac{k}{N(n)}$ for x from the set bounded by εC_k and εC_{k+1} .

Here, we suppose without loss of generality that the channels do not intersect the bottom and top faces of the square. The above function w^ε has been designed to possess the following properties :

- In the area situated between any two consecutive channels C_k and C_{k+1}, this function is equal to a constant, the constants are different in distinct areas.
- At each channel C_k the function w^ε makes a jump. The values of jumps are uniformly distributed on the channels so that the total increment of w^ε, as x_2 varies from 0 to 1, is equal to one.

By the definition and according to Proposition 3.3, the sequence w^ε is uniformly bounded in $W^{1,2}(Q_\varepsilon)$ and uniformly Lipschitz continuous; moreover, the Lipschitz constant is less than or equal to $c^{-1}(p)$. Thus, for a proper subsequence, we have

$$\begin{aligned} &w^\varepsilon \xrightarrow[\varepsilon\to 0]{} u_0 \quad \text{weakly in } W^{1,2}(Q_\varepsilon)\,,\\ &\sup_{x\in Q} |w^\varepsilon - u_0| \xrightarrow[\varepsilon\to 0]{} 0\,, \end{aligned}$$

where $u_0 \in W^{1,2}(Q)$, $u_0(\cdot, 0) = 0$, $u_0(\cdot, 1) = 1$, and

$$|u_0(x^1) - u_0(x^2)| \le c^{-1}(p)|x^1 - x^2|\,, \quad x^1,\, x^2 \in [0,1]^2\,.$$

Consider the expression

$$J^\varepsilon(w^\varepsilon) = \varepsilon^2 \sum_{x\in Q_\varepsilon}\sum_{z,z'\in\Lambda} a^\varepsilon_{zz'}(x)\,\partial^\varepsilon_z w^\varepsilon(x)\,\partial^\varepsilon_{z'} w^\varepsilon(x) = \varepsilon^2 \sum_{x\in Q_\varepsilon}\sum_{z\in\Lambda} p^\varepsilon_z(x)\,(\partial^\varepsilon_z w^\varepsilon(x))^2\ . \tag{3.8}$$

It follows from the definitions of w^ε and $p^\varepsilon_z(x)$ that

$$J^\varepsilon(w^\varepsilon) \le c^{-2}(p)\,\delta\,. \tag{3.9}$$

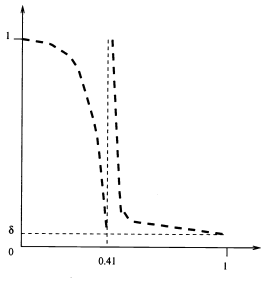


FIG. 3.3. *Illustration of Theorem* 3.10. *The behavior of* $a^\delta(p)$.

Moreover, by Proposition 2.14, $\liminf_{\varepsilon\to 0} J^\varepsilon(w^\varepsilon) \geq a^\delta(p) \int_{[0,1]^2} |\nabla u_0(x)|^2 dx \geq a^\delta(p)$. Combining the last two estimates, we get the desired inequality $a^\delta(p) \leq c^{-2}(p)\,\delta$. The lower bound $a^\delta(p) \geq \delta$ is evident. □

The next result describes the behavior of the effective coefficient $a^\delta(p)$ for p from a neighborhood of the critical point $p_c(\sqrt{2})$.

THEOREM 3.10. *In the vicinity of* $p_c(\sqrt{2})$, *the following inequalities hold:*

$$c_1\,(p_c(\sqrt{2}) - p)^{\alpha_1} \leq a^\delta(p) \quad \textit{if} \ \ p < p_c(\sqrt{2}),$$
$$a^\delta(p) \leq \frac{c_2}{(p - p_c(\sqrt{2}))^{\alpha_2}}\,\delta \quad \textit{if} \ \ p_c(\sqrt{2}) < p,$$

where c_1, c_2, α_1, *and* α_2 *are strictly positive constants.*

Figure 3.3 illustrates this result.

Proof. It is sufficient to substitute the estimate (3.1) in (3.7) and (3.9). The required estimates are now straightforward. □

3.2.2. Comparison with the behavior in continuous media. The asymptotic behavior of the effective coefficient described in the previous section (section 3.2.1) differs essentially from that obtained for the case of differential equations (see Jikov, Kozlov, and Oleinik [11, Chapter 9]). One of the reasons for this disagreement is the fact that we ignore the streams through the neighborhoods of vertices of the checker-board structure.

Here we modify the model of the previous section by involving the streams along the "diagonal directions," so that the asymptotic behavior of the effective coefficient as $\delta \to 0$ in this new model is similar to that obtained for the corresponding differential operator.

Let us begin by describing the scheme of discretization. We set

$$\Lambda = \{(0,0), \pm e_1, \pm e_2, \pm(e_1+e_2), \pm(e_1-e_2)\}, \qquad e_1 \overset{\triangle}{=} (1,0),\ e_2 \overset{\triangle}{=} (0,1),$$

(so, at each step, a trajectory of the corresponding random walk can choose one of the eight nearest points of $\mathbb{Z}^2$ or keep the same position).

In order to assign the values for $p_z(x)$, $|z| = \sqrt{2}$, we consider auxiliary periodic checker-board structure with a cell of periodicity shown in Figure 3.4. The effective coefficient of this medium is equal to $\sqrt{\delta}$ (see Jikov, Kozlov, and Oleinik [11, section 7.2]). This gives us an idea that, for the combination of squares shown in Figure 3.4, the coefficient $p_z(x)$ with $z = (e_1 + e_2)$, should be of order $\sqrt{\delta}$.

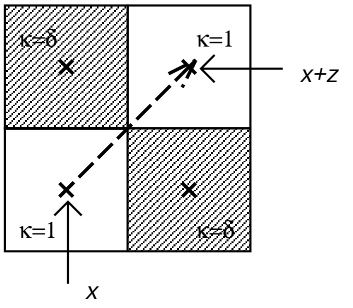


Fig. 3.4.

Inspired by these heuristic arguments, we define the transition probabilities by

$$p_z(x) = \begin{cases} \frac{1}{8}\min\left(\frac{2\,\kappa(x)\,\kappa(x+z)}{\kappa(x)+\kappa(x+z)},\ \sqrt{\frac{\kappa(x+z_1e_1)+\kappa(x+z_2e_2)}{2}\,\frac{\kappa(x)+\kappa(x+z)}{2}}\right) & \text{if } |z| = \sqrt{2}\,, \\ \dfrac{1}{8}\,\dfrac{2\,\kappa(x)\,\kappa(x+z)}{\kappa(x)+\kappa(x+z)} & \text{if } |z| = 1\,, \\ 1 - \sum_{z\in\Lambda, z\neq(0,0)} p_z(x) & \text{if } z = (0,0)\,, \\ p_z(x) = 0 & \text{if } z \notin \Lambda\,. \end{cases}$$

The following theorem describes the asymptotic behavior of the effective coefficient $a^\delta(p)$.

Theorem 3.11. *The effective coefficient $a^\delta(p)$ satisfies, for small δ, the estimates*

$$\begin{array}{ll} 0 < c_1(p) \le a^\delta(p) \le 1 & \textit{if}\ \ 0 \le p < p_c(\sqrt{2}), \\ c_2(p)\,\sqrt{\delta} \le a^\delta(p) \le c_3(p)\,\sqrt{\delta} & \textit{if}\ \ p_c(\sqrt{2}) < p < 1 - p_c(\sqrt{2}), \\ \delta \le a^\delta(p) \le c_4(p)\,\delta & \textit{if}\ \ 1 - p_c(\sqrt{2}) < p \le 1, \end{array}$$

where $c_1(p)$, $c_2(p)$, $c_3(p)$, and $c_4(p)$ are strictly positive.

Thus, the effective coefficient is uniformly positive when $p < p_c(\sqrt{2})$, is of order $\sqrt{\delta}$ when p is between $p_c(\sqrt{2})$ and $1 - p_c(\sqrt{2})$, and is of order δ when $p > 1 - p_c(\sqrt{2})$.

Proof. The cases $0 \le p < p_c(\sqrt{2})$ and $1 - p_c(\sqrt{2}) < p \le 1$ can be studied exactly in the same way as in Theorem 3.9. □

Now, we proceed with the case $p_c(\sqrt{2}) < p < 1 - p_c(\sqrt{2})$.

Consider the percolation model with $\gamma = \sqrt{2}$. Again, for sufficiently large n, there are at least $c(p)\,n$ mutually nonintersecting black $\sqrt{2}$-channels and white $\sqrt{2}$-channels joining the left and the right sides of the square $[0,n]^2$ (see Figure 3.5).

Lower bound. We consider the infinite white cluster. In order to obtain the lower bound for $a^\delta(p)$, we follow part (1) of the proof of Theorem 3.9. We point out that, along each white channel, if both x and $x+z$ belong to the channel and $|z| \le \sqrt{2}$, then $p_z(x) \ge \sqrt{\delta}/8$. Indeed, in this case $\kappa(x) = \kappa(x+z) = 1$ and, by the definition, $p_z(x)$ takes on one of the following values: $\frac{1}{8}$, $\frac{1}{8}\sqrt{\frac{1+\delta}{2}}$, $\frac{1}{8}\sqrt{\delta}$.

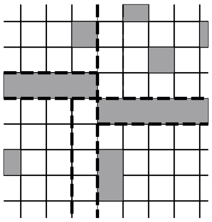


FIG. 3.5. *Intersection between a black and a white channel;* $p \in]\, p_c(\sqrt{2}),\, 1 - p_c(\sqrt{2})[$.

From (3.4), (3.6) and the above estimate of $p_z(x)$, we get

$$\begin{aligned} \|P_1(z) - \varphi\|^2 &= \lim_{n\to+\infty} \frac{1}{n^2} \sum_{x\in\mathbb{Z}^2\cap[0,n]^2} \sum_{z\in\Lambda(x)} p_z(T_x\,\omega)\, \Big(z_1 - \tilde{\phi}(T_{x+z}\,\omega) + \tilde{\phi}(T_x\,\omega)\Big)^2 \\ &\geq \liminf_{n\to+\infty} \frac{\sqrt{\delta}}{8\,n^2} \sum_{x\in C^w} \sum_{z\in\Lambda(x)} \Big(z_1 - \tilde{\phi}(T_{x+z}\,\omega) + \tilde{\phi}(T_x\,\omega)\Big)^2 \\ &\geq \lim_{n\to+\infty} \frac{\sqrt{\delta}}{8\,n^2} \frac{c(p)}{\theta(p)} (n-c)^2 \geq c\sqrt{\delta}\,, \end{aligned}$$

where symbol C^w stands for the union of white channels. By virtue of (3.3), the last inequality implies the required lower bound.

Upper bound. We consider the infinite black cluster and the $N(n) = c(p)\,n$ black channels C_k^b, $k = 1, 2, \dots, N(n)$ in the square $[0,\, n]^2$.

The upper bound $a^\delta(p) \leq c_3(p)\sqrt{\delta}$ can be established with the help of the following auxiliary functions:

- $w^\varepsilon(\cdot, 0) = 0\,,\ w^\varepsilon(\cdot, 1) = 1$;
- $w^\varepsilon(x) = \frac{(k-1/2)}{N(n)}$ for $x \in \varepsilon C_k^b$;
- $w^\varepsilon(x) = \frac{k}{N(n)}$ for x from the set bounded by εC_k^b and εC_{k+1}^b ,

where $\varepsilon = 1/n$. Direct calculations show that $J^\varepsilon(w^\varepsilon) \leq c^{-2}(p)\sqrt{\delta}$; indeed, by the definition of $\{p_z(x)\}$, we have $p_z(x) \leq \delta/8$ if x belongs to a black channel, and $p_z(x) \leq \frac{\sqrt{\delta}}{8}$ if x and $x+z$ are situated at the opposite banks of a black channel. If we denote by u_0 an accumulating point of w^ε, then we have by Proposition 2.14

$$c(p)^{-2}\,\sqrt{\delta} \geq \lim_{\varepsilon\to 0} J^\varepsilon(w^\varepsilon) \geq a^\delta(p) \int_{[0,\,1]^2} |\nabla u_0(x)|^2\, dx \geq a^\delta(p)\,.$$

Comparing these results with Jikov, Kozlov, and Oleinik [11, Chapter 9, Theorem 9.5] shows that the discrete operators considered in this section adopt the asymptotic properties of the corresponding differential operators.

3.2.3. Geometric mean.

We modify here the scheme of discretization of section 3.2.1 by taking the geometric mean in the definition of transition probabilities instead of the harmonic mean:

$$p_z(x) = \begin{cases} \dfrac{1}{4}\sqrt{\kappa(x)\,\kappa(x+z)} & \text{if } z \in \Lambda \setminus \{(0,0)\}\,, \\ 1 - \sum_{z\in\Lambda\setminus\{(0,0)\}} p_z(x) & \text{if } z = (0,0)\,, \\ 0 & \text{if } z \notin \Lambda\,, \end{cases}$$

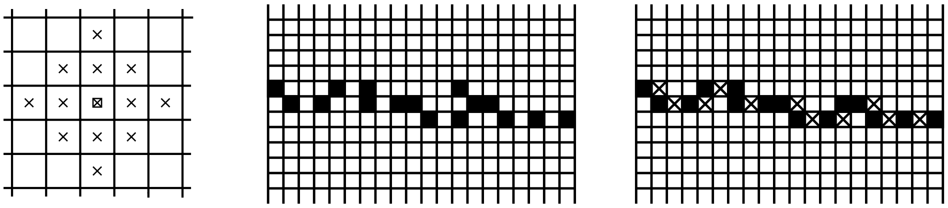


FIG. 3.6. $\gamma = 2$. *The neighbor squares (left), a black channel C_k (center), and one of its possible modifications $\tilde{C}_k$ (right).*

the set Λ being the same as in section 3.2.1 (i.e., with displacements toward the four nearest neighbors). Then, the asymptotic behavior of the effective coefficient $a^\delta(p)$ is described by the following statement.

THEOREM 3.12. *The effective coefficient $a^\delta(p)$ satisfies, for small δ, the estimates:*

$$\begin{array}{ll} 0 < c_1(p) \le a^\delta(p) \le 1 & \text{if } \ 0 \le p < p_c(\sqrt{2}), \\ c_2(p)\sqrt{\delta} \le a^\delta(p) \le c_3(p)\sqrt{\delta} & \text{if } \ p_c(\sqrt{2}) < p < 1 - p_c(2), \\ \delta \le a^\delta(p) \le c_4(p)\,\delta & \text{if } \ 1 - p_c(2) < p \le 1, \end{array}$$

where $c_1(p)$, $c_2(p)$, $c_3(p)$, and $c_4(p)$ are strictly positive.

Proof.

1. In the case $0 \le p < p_c(\sqrt{2})$, we need to justify only the lower bound. It can be done exactly in the same way as in Theorem 3.9. Another way to obtain the lower bound is to notice that for $|z| \ne 0$ the coefficients $p_z(x)$ under consideration majorate the respective coefficients defined as the harmonic mean. By virtue of the convergence of energy result and Theorem 3.9 this implies the desired lower bound.
2. In order to obtain the upper bound for $p_c(\sqrt{2}) < p < 1 - p_c(2)$ one can apply the technique developed in the part (2) of the proof of Theorem 3.9.

   To justify the lower bound in the case $p_c(\sqrt{2}) < p < 1 - p_c(2)$, we consider the percolation model with $\gamma = 2$ (see Remark 3.4). Here we encounter an additional difficulty: for $p \in]1 - p_c(\sqrt{2})\,, 1 - p_c(2)[$ the white 2-channels are not connected in a usual sense.

   We proceed as follows. For each channel C_k we introduce its 1-neighborhood:

   $$C_k^+ = \{x \in \mathbb{Z}^2 \,:\, |x - j| \le 1 \text{ for some } j \in C_k\}\,.$$

   It is easily seen that C_k^+ contains a sequence of squares $\{x_i\}$ denoted by $\tilde{C}_k$, which joins the left and the right sides of the square $[0, n]^2$ and has the following properties:

   - $|x_{i+1} - x_i| = 1$ for any consecutive x_i and x_{i+1};
   - $p_z(x) \ge \sqrt{\delta}/4$ for any x and z such that $x,\, z + z \in \tilde{C}_k$ and $|z| = 1$

   (see Figure 3.6) These sets $\tilde{C}_k$ are connected in a usual sense and consist in general of both white and black squares. Clearly, the number $\tilde{N}(n)$ of mutually nonintersecting sets $\tilde{C}_k$ still satisfies the estimate $\tilde{N}(n) \ge \tilde{c}(p)\,n$, $\tilde{c}(p) > 0$, for sufficiently large n. Then, one can use $\tilde{C}_k$ instead of C_k and argue like in part (1) of the proof of Theorem 3.9.
3. The upper bound in the case $1 - p_c(2) < p \le 1$ requires slightly different arguments than above. Consider the percolation model with $\gamma = 2$, and for each white cluster $\mathbb{C}$ denote by $\mathbb{C}^+$ the 1-neighborhood of $\mathbb{C}$:

   $$\mathbb{C}^+ = \{x \in \mathbb{Z}^2 \,:\, |x - j| \le 1 \text{ for some } j \in \mathbb{C}\}\,.$$

Let $\mathbb{C}^+(0)$ be the set $\mathbb{C}^+$ containing 0, and denote by $W(0)$ the size of $\mathbb{C}^+(0)$. If 0 does not belong to the 1-neighborhood of the union of white clusters, then $\mathbb{C}^+(0)$ is empty and $W(0) = 0$.

We introduce the following sequence of random variables $\tilde{\varphi}^N(\omega) \in L^\infty(\Omega)$:

$$\tilde{\varphi}^N = \begin{cases} -\min\limits_{j \in \mathbb{C}^+(0)} j_1 & \text{if } 1 \le W(0) \le N \\ 0 & \text{otherwise,} \end{cases}$$

and put $\varphi_z^N(\omega) = \tilde{\varphi}^N(T_z\omega) - \tilde{\varphi}^N(\omega)$, $z \in \Lambda$. It is clear that $|\varphi_z^N(\omega)| \le 2N$. According to Kesten [12, Theorem 5.1], the estimate

$$P_p\{W(0) > n\} \le c\,\exp(-c(p)\,n)\,, \quad c(p) > 0\,, \tag{3.10}$$

holds for all $p > 1 - p_c(2)$. Therefore, by the definition of φ_z^N, we have

$$P_p\{\varphi_z^N \ge n\} \le c\,\exp(-c_1(p)\,n)\,, \quad c_1(p) > 0\,, \quad n = 1, 2, \ldots, 2N\,. \tag{3.11}$$

The random variables φ_z^N and p_z possess the following properties:

- if both 0 and z belong to $\mathbb{C}^+(0)$ and $W(0) \le N$, then $P_1(z) - \varphi_z^N = 0$;
- if at least one of them does not belong to $\mathbb{C}^+(0)$, then $p_z = \delta/4$.

In combination with (3.10) and (3.11), this implies

$$\begin{aligned} a^\delta(p) \le \|P_1(z) - \varphi_z^N\| &= E\sum_{z\in\Lambda} p_z(z_1 - \varphi_z^N)^2 \\ &\le c\,\delta\sum_{k=1}^{2N} k\,\exp(-c_1(p)k) + c\,\exp(-c(p)N) \\ &\le \bar{c}\,\delta + c\,\exp(-c(p)N)\,, \end{aligned}$$

where $\bar{c}$ does not depend on N. Passing to the limit as $N \to \infty$ gives $a^\delta(p) \le \bar{c}\,\delta$. □

3.2.4. Arithmetic mean. This section deals with another modification of the scheme of section 3.2.1. Namely, the transition probabilities are defined as the corresponding arithmetic means

$$p_z(x) = \begin{cases} \dfrac{1}{4}\dfrac{\kappa(x) + \kappa(x+z)}{2} & \text{if } z \in \Lambda \setminus \{(0,0)\}\,, \\ 1 - \sum_{z\in\Lambda\setminus\{(0,0)\}} p_z(x) & \text{if } z = (0,0)\,, \\ 0 & \text{if } z \notin \Lambda\,, \end{cases}$$

while the set Λ remains the same as in section 3.2.1.

THEOREM 3.13. *The effective coefficient $a^\delta(p)$ satisfies, for small δ, the estimates*

$$\begin{aligned} &0 < c_1(p) \le a^\delta(p) \le 1 \quad && \textit{if } \ 0 \le p < 1 - p_c(2), \\ &\delta \le a^\delta(p) \le c_2(p)\,\delta && \textit{if } \ 1 - p_c(2) < p \le 1, \end{aligned}$$

where $c_1(p)$ and $c_2(p)$ are strictly positive.

Proof. The first estimate relies on the channel property of the percolation model corresponding to $\gamma = 2$. As in the preceding theorem, we enlarge the white 2-channels to make them connected, and note that along each modified channel the transition probabilities are uniformly positive: $p_z(x) \ge (1+\delta)/8$ if $z \in \Lambda$ and x and $x+z$ belong to a modified channel. As above, this implies the lower bound $a^\delta(p) \ge c_1(p) > 0$.

The proof of the second estimate is exactly the same as that of the last estimate in the preceding theorem. □

Remark 3.14. *The statements of Theorems* 3.9–3.11 *remain unchanged if we assume that the size of mesh $h(\varepsilon)$ of a grid is less than ε while $h(\varepsilon)/\varepsilon$ is a constant.*

Appendices.

Appendix A. Convergence of discrete functions. Let f^ε be an arbitrary function defined in the discrete domain $Q_\varepsilon = \varepsilon\mathbb{Z}^d \cap Q$, and let $\tilde{f}^\varepsilon$ be the piecewise-constant interpolation of f^ε:

$$\tilde{f}^\varepsilon(x) = f^\varepsilon(y) \quad \text{if } y \in Q_\varepsilon \text{ and } x \in y + \left[\frac{-\varepsilon}{2}, \frac{\varepsilon}{2}\right]^d .$$

Definition A.1. *We say that a family of functions $f^\varepsilon \in L^2(Q_\varepsilon)$ converges strongly (resp., weakly) to the function $f \in L^2(Q)$ as $\varepsilon \to 0$ if $\tilde{f}^\varepsilon$ converges strongly (resp., weakly) to f in $L^2(Q)$. For this convergence we use the notation*

$$f^\varepsilon \xrightarrow[\varepsilon\to 0]{} f \quad \text{in } L^2(Q_\varepsilon) \quad (\text{resp., weakly in } L^2(Q_\varepsilon)) .$$

Similarly, one can define the $W^{1,2}(Q)$-convergence of discrete functions with $\tilde{f}^\varepsilon$ being the piecewise linear interpolation of f^ε (instead of the piecewise constant one).

The convergence in $W^{-1,2}(Q)$ can be defined in terms of duality. Namely, we say that $f^\varepsilon \in W^{-1,2}(Q_\varepsilon)$ converges to $f \in W^{-1,2}(Q)$ strongly (resp., weakly) if for any sequence $g^\varepsilon \in W_0^{1,2}(Q_\varepsilon)$ and $g \in W_0^{1,2}(Q)$ such that $g^\varepsilon \to g$ weakly (resp., strongly) in $W^{1,2}(Q)$, we have

$$\langle f^\varepsilon, g^\varepsilon\rangle \xrightarrow[\varepsilon\to 0]{} \langle f, g\rangle .$$

Definition A.2. *Let $w^\varepsilon \in L^2(Q_\varepsilon)$ and $w^0 \in L^2(Q)$. The sequence w^ε* converges $\star$-weakly *to w^0 if for any $\varphi \in \mathcal{C}_0^\infty(Q)$,*

$$\lim_{\varepsilon\to 0} \varepsilon^d \sum_{x\in Q_\varepsilon} w^\varepsilon(x)\,\varphi(x) = \int_Q w^0(x)\,\varphi(x)\,dx .$$

Appendix B. The derivative of a product of discrete functions.

Proposition B.1. *Let f and g belong to $W^{1,2}(Q_\varepsilon)$. Then,*

$$\sum_{z\in Q_\varepsilon} |\partial_z^\varepsilon(fg) - f\partial_z^\varepsilon g - g\partial_z^\varepsilon f| \le \varepsilon \sum_{z\in Q_\varepsilon} |\partial_z^\varepsilon f||\partial_z^\varepsilon g| .$$

Proof. We have

$$\begin{aligned}\varepsilon\,\partial_z^\varepsilon(f(x)\,g(x)) &= f(x+\varepsilon\,z)\,g(x+\varepsilon\,z) - f(x)\,g(x)\\ &= g(x)\,(f(x+\varepsilon\,z) - f(x)) + f(x+\varepsilon\,z)\,(g(x+\varepsilon\,z) - g(x))\\ &= \varepsilon\,[g(x)\,\partial_z^\varepsilon f(x) + f(x+\varepsilon\,z)\,\partial_z^\varepsilon g(x)] .\end{aligned}$$

We have $f(x+\varepsilon\,z) = f(x) + \varepsilon\partial_z^\varepsilon f(x)$. Therefore,

$$\partial_z^\varepsilon(f(x)\,g(x)) = g(x)\,\partial_z^\varepsilon f(x) + f(x)\,\partial_z^\varepsilon g(x) + \varepsilon\partial_z^\varepsilon f(x)\partial_z^\varepsilon g(x),$$

and the desired estimate immediately follows. □

Appendix C. The Friedrichs and Poincaré inequalities.

This appendix is devoted to the Friedrichs and Poincaré inequalities for grid functions. In fact, in order to prove the propositions below, one can follow the same ideas as in the case of the continuous argument. For this reason, we omit the proof.

PROPOSITION C.1. *Let Q be a bounded domain with piecewise smooth boundary and denote the discretization of Q by Q_ε. Then, for any $v^\varepsilon \in W_0^{1,2}(Q_\varepsilon)$ the following inequality holds:*

$$\|v^\varepsilon\|^2_{L^2(Q_\varepsilon)} \leq c(Q)\,\varepsilon^d \sum_{x\in Q_\varepsilon}\sum_{i=1}^{d}(\partial^\varepsilon_{\pm e_i} v^\varepsilon(x))^2 \,. \tag{C.1}$$

PROPOSITION C.2. *Let Q be a smooth bounded domain. Then, for all sufficiently small ε and for any $v^\varepsilon \in W^{1,2}(Q_\varepsilon)$ such that $\sum_{x\in Q_\varepsilon} v^\varepsilon(x) = 0$, the following inequality is satisfied:*

$$\sum_{x\in Q_\varepsilon} |v^\varepsilon(x)|^2 \leq C(q)\,\varepsilon^d \sum_{x\in Q_\varepsilon}\sum_{i=1}^{d} |\bar\partial^\varepsilon_{\pm e_i} v^\varepsilon(x)|^2 \,. \tag{C.2}$$

REMARK C.3. *The statement of Proposition C.2 remains valid for the domain $\overline{Q}_\varepsilon$.*

REFERENCES

[1] M. AIZENMAN AND G. R. GRIMMETT, *Strict monotonicity for critical points in percolation and ferromagnetic models*, J. Statist. Phys., 63 (1991), pp. 817–835.

[2] M. AVELLANEDA, TH. Y. HOU, AND G. C. PAPANICOLAOU, *Finite difference approximations for partial differential equations with rapidly oscillating coefficients*, M2AN Math. Model. Numer. Anal., 25 (1991), pp. 693–710.

[3] N. S. BAKHVALOV AND G. S. PANASENKO, *Homogenization of Processes in Periodic Media*, Nauka, Moscow, 1984.

[4] A. BENSOUSSAN, J. L. LIONS, AND G. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, Stud. Math. Appl. 5, North Holland, Amsterdam, 1978.

[5] L. BERLYAND AND K. GOLDEN, *Percolation analysis for effective conductivity of a continuum percolation model*, Phys. Rev. B, 50 (1994), pp. 2114–2117.

[6] J. BRICMONT AND A. KUPIAINEN, *Random walks in asymmetric random environments*, Comm. Math. Phys. 142 (1991), pp. 345–420.

[7] E. DE GIORGI, *Sulla convergenza di alcune successioni di integrali del tipo dell'area*, Rend. Mat. (6), 8 (1975), pp. 277–294.

[8] E. DE GIORGI, *G-operators and Γ-convergence*, in Proceedings of the International Congress Math., Warszawa, 1983, PWN Polish Scientific Publishers and North-Holland, 1984, pp. 1175–1191.

[9] K. GOLDEN AND S. M. KOZLOV, *Percolation Analysis of Effective Permeability of Porous Media*, Rapport 94-22, Laboratoire d'Analyse, Topologie et Probabilité, Université de Provence, 1994.

[10] G. R. GRIMMETT, *Percolation*, Springer-Verlag, New York, 1989.

[11] V. V. JIKOV, S. M. KOZLOV, AND O. A.OLEINIK, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, New York, 1994.

[12] H. KESTEN, *Percolation theory for Mathematicians*, in Progress in Probability and Statistics, Vol. 2, Birkhäuser, Boston, 1982.

[13] S. M. KOZLOV, *Averaging differential operators with almost periodic, rapidly oscillating coefficients*, Math. USSR-Sb., 35 (1979), pp. 481–498.

[14] S. M. KOZLOV, *The method of averaging and walks in inhomogeneous environments*, Russian Math. Surveys, 40 (1985), pp. 73–145.

[15] S. M. Kozlov, *Averaging of difference schemes*, Math. USSR Sb, 57 (1987), pp. 351–369.

[16] M. B. Krasniansky, *Homogenization of random walks on lattices with weak connections*, Math. Phys. Anal. Geom., 1 (1995), pp. 51–67 (in Russian).

[17] R. Künnemann, *The diffusion limit for reversible jump processes on $\mathbb{Z}^d$ with ergodic random bond conductivities*, Comm. Math. Phys., 90 (1983), pp. 27–68.

[18] G. Dal Maso, *An Introduction to Γ-convergence*, Birkhäuser, Boston, 1993.

[19] J. F.McCarthy, *Effective permeability of sandstone–shale reservoirs by random walk method*, J. Phys. A, 23 (1990), pp. 445–451.

[20] N. G. Meyers, *An L^p–estimate for the gradient of solutions of second order elliptic divergent equations*, Ann. Scuola Norm. Sup. Pisa (3), 17 (1963), pp. 189–206.

[21] F. Murat, *Compacité par compensation*, Ann. Scuola Norm. Sup Pisa. 4, 5 (1978), pp. 489–507.

[22] F. Murat and L. Tartar, *H-convergence*, in Topics in the Mathematical Modelling of Composite Materials, Progr. Nonlinear Differential Equations Appl. 31, Birkhäuser, Boston, MA, 1997.

[23] B. Nœtinger, *The effective permeability of heterogeneous porous media*, Transp. Porous Media, 15 (1994), pp. 99–127.

[24] G. C. Papanicolaou and S. R. S. Varadhan, *Diffusions with random coefficients*, in Statistics and Probability: An Essay in Honor of C. R. Rao, G. Kallianpur, P. R. Krishnaiah, and J. K. Ghosh, eds., North-Holland, Amsterdam, New York, 1982, pp. 547–552.

[25] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer-Verlag, New York, 1994.

[26] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, Vol. IV, Res. Notes in Math. 39, Pitman, Boston, London, 1979, pp. 136–211.

[27] S. Spagnolo, *Sulla convergenza delle soluzioni di equazioni parabolich ed ellitiche*, Ann. Scuola Norm. Sup. Pisa Cl. Sci, 22 (1968), pp. 571–597.

[28] V. V. Zhikov, S. M. Kozlov, O. A.Oleinik, and Kha T'en Ngoan, *Averaging and G-convergence of differential operators*, Russian Math. Surveys 34 (1979), pp. 69–147.

SIAM J. MATH. ANAL.
Vol. 33, No. 1, pp. 84–112

ABOUT LIFESPAN OF REGULAR SOLUTIONS OF EQUATIONS RELATED TO VISCOELASTIC FLUIDS[*]

JEAN-YVES CHEMIN[†] AND NADER MASMOUDI[‡]

Abstract. We prove existence and uniqueness of local and global solutions for a system of equations concerning an incompressible viscoelastic fluid of the Oldroyd type. We also show a new a priori estimate for the two-dimensional Navier–Stokes system and a losing estimate for the transport equations that allow us to give a sufficient condition of non-breakdown.

Key words. local and global well-posedness, Besov spaces, Oldroyd model

AMS subject classifications. 76A05, 76D03

PII. S0036141099359317

1. Introduction and statement of the results.

An incompressible fluid is subject to the following system of equations:

$$\left\{ \begin{array}{rcl} \dfrac{\partial v}{\partial t} + v \cdot \nabla v & = & \nabla \cdot \sigma, \\ \operatorname{div} v & = & 0, \end{array} \right.$$

where v is the velocity ($v(x,t) \in \mathbb{R}^d$) and σ is the stress tensor (σ is a (d,d) symmetric matrix). Moreover, σ can be decomposed as $\sigma = \tau - p\, Id$, where τ is the tangential part of the stress tensor and $-p\, Id$ is the normal part (p being the pressure which is the Lagrange multiplier for the divergence-free condition). For a Newtonian fluid, τ depends linearly on ∇v and more precisely

$$\tau = 2\nu D(v), \tag{1.1}$$

where $D(v) = \frac{1}{2}(\nabla v +^t \nabla v)$ is the deformation tensor and ν is the viscosity of the fluid ($\nu > 0$). Hence, we recover the classical incompressible Navier–Stokes system.

It turns out that many fluids do not satisfy the Newtonian law $\tau = 2\nu D(v)$ and a general constitutive law satisfied by all fluids does not exist. Some fluids with shear dependent viscosity are such that (1.1) is replaced by

$$\tau = 2\mu(|D(v)|^2) D(v), \tag{1.2}$$

where the viscosity $\mu(|D(v)|^2)$ depends on $|D(v)|$. In the case of the power law, we have

$$\mu(|D(v)|^2) = \nu + \beta |D(v)|^{p-2},$$

[*] Received by the editors July 19, 1999; accepted for publication (in revised form) November 9, 2000; published electronically April 24, 2001.
http://www.siam.org/journals/sima/33-1/35931.html

[†] Analyse Numérique, Case 187, Université Pierre et Marie CURIE, 4 Place Jussieu, 75230 Paris Cedex 05, France (chemin@ann.jussieu.fr).

[‡] Ceremade-URA CNRS 749, Université Paris IX Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 PARIS Cedex 16, France (masmoudi@dmi.ens.fr).

where $\nu \geq 0$, $\beta > 0$, and $p \geq 1$. If

$1 \leq p < 2$, we have a shear thinning fluid or a viscoplastic fluid,

$p = 2$, we have the classical Newtonian case,

$p > 2$, we have a shear thickening fluid or dilatant fluid.

In this paper, we are going to study another type of non-Newtonian fluid, namely, fluids with memory. Indeed, for many fluids, it is not possible to determine at some time t the value of τ knowing only $D(u)$ at the same time, and one also has to know the whole history of $D(v)$. In these cases, we say that the fluid has a "memory" and one has to write a differential equation for τ. One of the classical models is the Oldroyd model, which writes

$$\partial_t \tau + v \cdot \nabla \tau + a\tau + F(\tau, \nabla v) = 0,$$

where $a > 0$ and F is a quadratic form in $(\tau, \nabla v)$. It turns out that since the model should be invariant under change of coordinates, F cannot be the most general quadratic form and we get the so-called Oldroyd model with eight constants, which can be rewritten in the following way:

$$\tau + \lambda_1 \frac{\mathcal{D}\tau}{\mathcal{D}t} + \frac{1}{2}\mu_0 tr(\tau) D(v) - \frac{1}{2}\mu_1 \{\tau D(v) + D(v)\tau\} + \frac{1}{2}\nu_1 (\tau : D(v)) Id$$

$$= 2\eta_0 \left[D(v) + \lambda_2 \frac{\mathcal{D}D(v)}{\mathcal{D}t} - \mu_2 D(v)D(v) + \frac{1}{2}(D(v) : D(v)) Id \right],$$

where we have used "objective derivatives" (the so-called Oldroyd derivatives)

$$\frac{\mathcal{D}\tau}{\mathcal{D}t} = \frac{\partial \tau}{\partial t} + v \cdot \nabla \tau + \tau W(v) - W(v)\tau,$$

and $W(v) = \frac{1}{2}(\nabla v -^t \nabla v)$ is the vorticity tensor.

In this paper, we are going to study a simpler model, namely the Oldroyd B model (with only four constants). It is given by

$$\tau + \lambda_1 \frac{\mathcal{D}_b \tau}{\mathcal{D}t} = 2\eta \left(D(v) + \lambda_2 \frac{\mathcal{D}_b D(v)}{\mathcal{D}t} \right), \tag{1.3}$$

where

$$\frac{\mathcal{D}_b \tau}{\mathcal{D}t} = \frac{\partial \tau}{\partial t} + v \cdot \nabla \tau + \tau W(v) - W(v)\tau - b(D(v)\tau + \tau D(v)).$$

In (1.3), λ_1 is the relaxation time, λ_2 is the retardation time ($0 \leq \lambda_2 \leq \lambda_1$), η is the dynamical viscosity of the fluid, and $b \in [-1, 1]$. Fluids of this type have both elastic properties and viscous properties. Indeed, the case $\lambda_2 = \lambda_1 = 0$ corresponds to purely viscous case (incompressible Navier–Stokes equation), while the case $\lambda_1 > \lambda_2 = 0$ is the purely elastic case (the Maxwell model). Decomposing τ into

$$\tau = \tau_{Newtonian} + \tau_{elastic} \quad \text{with} \quad \tau_{Newtonian} = 2\eta \frac{\lambda_2}{\lambda_1} D(v)$$

we find that $\tau_{elastic}$ satisfies

$$\tau_{elastic} + \lambda_1 \frac{\mathcal{D}_b \tau_{elastic}}{\mathcal{D}t} = 2\eta \left(1 - \frac{\lambda_2}{\lambda_1}\right) D(v).$$

Taking

$$a = \frac{1}{\lambda_1}, \quad \mu_2 = \frac{2\eta}{\lambda_1}\left(1 - \frac{\lambda_2}{\lambda_1}\right), \quad \text{and} \quad \mu_1 = 1$$

and writing τ instead of $\tau_{elastic}$, we get the following system of equations:

$$(1.4) \quad (VE) \begin{cases} \dfrac{\partial v}{\partial t} + v\cdot\nabla v - \nu\Delta v + \nabla p &= \mu_1 \nabla\cdot\tau \quad \text{in} \quad \Omega\times(0,T), \\ \dfrac{\partial \tau}{\partial t} + v\cdot\nabla\tau + a\tau + Q(\tau,\nabla v) &= \mu_2 D(v) \quad \text{in} \quad \Omega\times(0,T), \\ \text{div } v &= 0 \quad \text{in} \quad \Omega\times(0,T), \end{cases}$$

where $\Omega = \mathbf{R^d}$ or $\Omega = \mathbf{T^d}$, u is the velocity vector field ($v(x,t) \in \mathbf{R^d}$), τ is the non-Newtonian part of the stress tensor, ($\tau(x,t)$ is a (d,d) symmetric matrix), $(\nabla\cdot\tau)_i = \sum_j \partial_j \tau_{i,j}$, and p is the pressure which is a scalar. The constants ν, a, μ_1, μ_2 are assumed to be nonnegative and the bilinear term Q has the following form:

$$Q(\tau,\nabla v) = \tau W(v) - W(v)\tau - b(D(v)\tau + \tau D(v)),$$

$b \in [-1,1]$, $D(v) = \frac{1}{2}(\nabla v +^t \nabla v)$ is the deformation tensor, and $W(v) = \frac{1}{2}(\nabla v -^t \nabla v)$ is the vorticity tensor. The system must be complemented with the following initial conditions:

$$(1.5) \qquad \begin{cases} v(0,\cdot) = v_0 & \text{in} \quad \Omega, \\ \tau(0,\cdot) = \tau_0 & \text{in} \quad \Omega. \end{cases}$$

Throughout this paper, solution means solution in the sense of distributions. As usual, problems of regularity are motivated by the uniqueness problem.

The formal energy estimate is the following:

$$\frac{1}{2}\frac{d}{dt}(\mu_2\|v(t)\|_{L^2}^2 + \mu_1\|\tau(t)\|_{L^2}^2) + \nu\mu_2\|\nabla v(t)\|_{L^2}^2 + a\mu_1\|\tau(t)\|_{L^2}^2 \leq |b|\,\|Dv(t)\|_{L^\infty}\|\tau(t)\|_{L^2}^2.$$

The system (1.4) describes the motion of an incompressible fluid satisfying the Oldroyd [19] constitutive law. The existence and uniqueness of local strong solutions in Hilbert spaces H^s have been established by Guillopé and Saut in [14]. These solutions are global if the coupling between the two equations is small as well as the initial data [15]. The case of L^s–L^r solutions has been treated by Fernandez Cara, Guillén, and Ortega in [16]. Results for the stationary problem are due to M. Renardy (see [20]). Recently, for $b = 0$, the existence of global weak solutions has been proved by Lions and Masmoudi [17].

In this paper, we show existence and uniqueness results for local and global solutions in some limit spaces, i.e., spaces invariant by the Navier–Stokes scaling. We also show that in two dimensions, the $L^1_T(L^\infty_x)$ norm of τ controls the equation. For this we show two results about the two-dimensional (2-D) Navier–Stokes system and about a losing a priori estimate for the transport equation satisfied by τ.

Our first result is related to Sobolev spaces and uses the special structure of the system (VE).

Theorem 1.1. *Let (v_0, τ_0) be an initial data in H^s with s strictly greater than $d/2$. Then a unique strictly positive maximal time $T^\star$ exists so that a unique solution (v, τ) exists in the space $L^\infty_{loc}([0, T^\star[; H^s)$. Moreover, this solution is such that v belongs to*

$$L^2_{loc}([0, T^\star[; H^{s+1}) \cap L^\infty_{loc}(]0, T^\star[; H^{s+1-\varepsilon}) \quad \textit{for any strictly positive } \varepsilon.$$

We have the following necessary condition for blow up:

$$T^\star < \infty \Longrightarrow \int_0^{T^\star} \left(\frac{\mu_1}{\mu_2 \nu} \|\tau(t)\|^2_{L^\infty} + \|\nabla v(t)\|_{L^\infty} \right) dt = +\infty.$$

As we shall see in section 2, the proof of this result in very classical. It is in the spirit of the well-known Beale–Kato–Majda criterion (see [2]). We want to improve the above necessary condition for blow up. The theorem is the following.

Theorem 1.2. *In two space dimensions, the necessary condition for blow up of Theorem 1.1 above becomes*

$$T^\star < \infty \Longrightarrow \int_0^{T^\star} (\|\tau(t)\|_{L^\infty} + |b| \, \|\tau(t)\|^2_{L^2}) dt = +\infty.$$

The first thing to notice here is that the required regularity is far from the regularity prescribed by the scaling. Let us say a word about this. One of the key concepts of the fundamental work of Fujita and Kato (see [12]) about local well-posedness for the incompressible Navier–Stokes system is the scaling invariance. It means that if a vector field v is a solution of incompressible Navier–Stokes system with initial data v_0, then $v_\lambda(t, x) \stackrel{\text{def}}{=} \lambda v(\lambda^2 t, \lambda x)$ is a solution of incompressible Navier–Stokes system with initial data $v_{0,\lambda}(x) \stackrel{\text{def}}{=} \lambda v_0(\lambda x)$. An easy computation will convince the reader that

$$|v_{0,\lambda}|_{H^{\frac{d}{2}-1}} = |v_0|_{H^{\frac{d}{2}-1}}.$$

We want to solve the system (VE) for initial data whose regularity fits with this scaling, as, for instance, for the usual incompressible Navier–Stokes system. This requires the use of Besov spaces. Let us recall the definitions of these spaces. For this, we need the Littlewood–Paley decomposition.

Proposition 1.3. *Let us denote by $\mathcal{D}(\Omega)$ the space of C^∞ functions whose support is compact and included in Ω. Let us define $\mathcal{C}$ to be the ring of center 0 of small radius $1/2$ and great radius 2. There exist two nonnegative radial functions χ and φ belonging, respectively, to $\mathcal{D}(B(0,1))$ and to $\mathcal{D}(\mathcal{C})$ so that*

$$\chi(\xi) + \sum_{q \geq 0} \varphi(2^{-q}\xi) = 1, \tag{1.6}$$

$$|p - q| \geq 2 \Rightarrow \mathrm{Supp}\ \varphi(2^{-\mathrm{q}}\cdot) \cap \mathrm{Supp}\ \varphi(2^{-\mathrm{p}}\cdot) = \emptyset. \tag{1.7}$$

For instance, one can take $\chi \in \mathcal{D}(B(0,1))$ such that $\chi \equiv 1$ on $B(0, 1/2)$ and take

$$\varphi(\xi) = \chi(2\xi) - \chi(\xi).$$

Then we are able to define the Littlewood–Paley decomposition. Let us denote by $\mathcal{F}$ the Fourier transform on $\mathbf{R^d}$. Let $h,\ \widetilde{h},\ \Delta_q,\ S_q\ (q \in \mathbb{Z})$ be defined as follows:

$$
\begin{aligned}
h = \mathcal{F}^{-1}\varphi \quad \text{and} \quad \widetilde{h} &= \mathcal{F}^{-1}\chi, \\
\Delta_q u = \mathcal{F}^{-1}(\varphi(2^{-q}\xi)\mathcal{F}u) &= 2^{qd}\int h(2^q y)u(x-y)dy, \\
S_q u = \mathcal{F}^{-1}(\chi(2^{-q}\xi)\mathcal{F}u) &= 2^{qd}\int \widetilde{h}(2^q y)u(x-y)dy.
\end{aligned}
$$

In this paper, we shall use Bony's decomposition which consists of writing

$$
uv = T_u v \ + \ T_v u \ + \ R(u,v),
$$

where

$$
T_u v \overset{\text{def}}{=} \sum_{q \in \mathbf{Z}} S_{q-1}u\Delta_q v \quad \text{and} \quad R(u,v) \overset{\text{def}}{=} \sum_{|q-q'|\leq 1} \Delta_{q'}u\Delta_q v.
$$

The Besov spaces we are going to use are homogeneous ones.

Definition 1.4. *Let s be a real number and let p and r be two real numbers greater than 1. Then we define the norm*

$$
\|u\|_{\widetilde{B}^s_{p,r}} \overset{def}{=} \|S_0 u\|_{L^p} + \left\| (2^{qs}\|\Delta_q u\|_{L^p})_{q\in\mathbf{N}} \right\|_{\ell^r(\mathbf{N})}
$$

and the seminorm

$$
\|u\|_{B^s_{p,r}} \overset{def}{=} \left\| (2^{qs}\|\Delta_q u\|_{L^p})_{q\in\mathbf{Z}} \right\|_{\ell^r(\mathbf{Z})}.
$$

Definition 1.5.

- *Let s be a real number and let p and r be two real numbers greater than 1. We denote by $\widetilde{B}^s_{p,r}$ the space of tempered distributions u such that $\|u\|_{\widetilde{B}^s_{p,r}}$ is finite.*
- *If $s < d/p$ or $s = d/p$ and $r = 1$, we define the homogeneous Besov space $B^s_{p,r}$ as the closure of compactly supported smooth functions for the norm $\|\cdot\|_{B^s_{p,r}}$.*

Remarks. It is obvious that all of those spaces are Banach spaces. Moreover, if $p = 2$, the Besov spaces can be described in the following way. The norm $\|u\|_{B^s_{2,r}}$ is equivalent to

$$
\left\| 2^{qs}\left(\int_{2^q \leq |\xi| \leq 2^{q+1}} |\widehat{u}(\xi)|^2 d\xi \right)^{\frac{1}{2}} \right\|_{\ell^r(\mathbf{Z})}.
$$

Let us point out that $B^s_{2,2}$ is a usual Sobolev space H^s and that $B^s_{\infty,\infty}$ is the usual Hölder space C^s.

Now we state the following theorems.

Theorem 1.6 (existence). *Let p be in $[1,+\infty[$ and let us define $s_c \overset{def}{=} \frac{d}{p}$.*

If (v_0,τ_0) belongs to $B^{s_c-1}_{p,1} \times B^{s_c}_{p,1}$, then a strictly positive real number T exists such that a solution (v,τ) of (VE) exists on $[0,T]\times\mathbf{R^d}$; this solution belongs to

$$
C([0,T];B^{s_c-1}_{p,1}) \cap L^\infty_{loc}(]0,T];B^{s_c}_{p,\infty}) \cap L^1([0,T];B^{s_c+1}_{p,1}) \times L^\infty([0,T];B^{s_c}_{p,1}).
$$

Moreover, a strictly positive constant c exists such that if

$$\mu_1\mu_2 \leq c\ a\nu \quad \textit{and} \quad \|v_0\|_{B^{s_c-1}_{p,1}} + \frac{\mu_1}{a}\|\tau_0\|_{B^{s_c}_{p,1}} \leq c\nu,$$

then the above time T may be $+\infty$.

Moreover, if we define T_{VE} by

$$T_{VE} \overset{def}{=} +\infty \;\; \textit{if}\; \mu_1\mu_2 \leq c\nu a \quad \textit{and} \quad T_{VE} \overset{def}{=} -\frac{1}{a}\log\left(1 - \frac{c\nu a}{2\mu_1\mu_2}\right) \;\; \textit{if}\; \mu_1\mu_2 \geq c\nu a, \tag{1.8}$$

then the above time T may be bounded from below by $\min\{T_{VE}, T_{(v_0,\tau_0)}\}$, where $T_{(v_0,\tau_0)}$ is the greatest strictly positive number such that

$$\mu_1\frac{1-e^{-aT}}{a}\|\tau_0\|_{B^{s_c}_{p,1}} + \sum_q 2^{q(s_c-1)}\|\Delta_q v_0\|_{L^p}(1 - e^{-c\nu 2^{2q}T}) \leq c\nu.$$

Now let us state a uniqueness theorem. The uniqueness, as, for instance, in the work of Danchin and Desjardins about the KdV-type model (see [10]), requires a restriction on p.

Theorem 1.7 (uniqueness). *Let p be in $[1, 2d[$ and let us define $s_c \overset{def}{=} \frac{d}{p}\cdot$*

If (v_0, τ_0) belongs to $B^{s_c-1}_{p,1} \times B^{s_c}_{p,1}$, then a unique strictly positive real number $T^\star$ exists such that a unique solution (v, τ) of (VE) exists on $[0, T^\star[\times \mathbf{R^d}$ in the space

$$C([0, T^\star[; B^{s_c-1}_{p,1}) \cap L^1_{loc}([0, T^\star[; B^{s_c+1}_{p,1}) \times L^\infty_{loc}([0, T^\star[; B^{s_c}_{p,1}).$$

Moreover, a strictly positive constant c exists such that, if

$$\mu_1\mu_2 \leq c\ a\nu \quad \textit{and} \quad \|v_0\|_{B^{s_c-1}_{p,1}} + \frac{\mu_1}{a}\|\tau_0\|_{B^{s_c}_{p,1}} \leq c\nu,$$

then the above time T may be $+\infty$. And if $T^\star$ is finite, then

$$\int_0^{T^\star} (\|\nabla v(t,\cdot)\|_{L^\infty} + \|\tau(t,\cdot)\|_{L^\infty})dt = +\infty.$$

Now we want to state an equivalent of Theorem 1.2 in the framework of Besov spaces. Unfortunately, it will be impossible to get it for critical regularity. Thus, we need the following theorem.

Theorem 1.8. *Let p be in $[1, 2d[$ and let us assume that, for some σ strictly greater than s_c, the initial data (v_0, τ_0) belongs to $B^{s_c-1}_{p,1} \cap B^{\sigma-1}_{p,1} \times B^{s_c}_{p,1} \cap B^{\sigma}_{p,1}$, then the (unique) solution belongs to*

$$C([0, T^\star[; B^{\sigma-1}_{p,1}) \cap L^\infty_{loc}(]0, T^\star[; B^{\sigma}_{p,\infty}) \cap L^1_{loc}([0, T^\star[; B^{\sigma+1}_{p,1}) \times L^\infty_{loc}([0, T^\star[; B^{\sigma}_{p,1}).$$

Moreover, if the initial data (v_0, τ_0) belongs to L^2, and if $b = 0$, then the solution satisfies the energy estimate

$$\mu_2\|v(t)\|^2_{L^2} + \mu_1\|\tau(t)\|^2_{L^2} + 2\int_0^t \Big(\nu\mu_2\|\nabla v(t')\|_{L^2} + a\mu_1\|\tau(t')\|^2_{L^2}\Big)dt' = \mu_2\|v_0\|^2_{L^2} + \mu_1\|\tau_0\|^2_{L^2}.$$

Finally, for this last theorem, we take $p = 2$, $s \geq d/2$ and impose the same regularity for v_0 and τ_0. In fact, we are going to use the particular form of the equation and will not have to impose any restriction on the coupling to get the global existence. For this theorem, we take nonhomogenous Besov spaces. We can also work in Sobolev spaces if we impose in addition that $s > d/2$.

Theorem 1.9. *Let $s \geq \frac{d}{2}$. A constant c exists such that for all nonnegative constants ν, a, μ_1, μ_2 and for any $(v_0, \tau_0) \in \widetilde{B}^s_{2,1} \times \widetilde{B}^s_{2,1}$ satisfying*

$$\|v_0\|_{B^s_{2,1}} \leq c \ \min(\nu, \sqrt{a\nu}) \quad \textit{and} \quad \|\tau_0\|_{B^s_{2,1}} \leq c \ \min(\nu, \sqrt{a\nu}) \frac{\sqrt{\mu}_2}{\sqrt{\mu}_1},$$

a unique global solution (v, τ) of (1.4) exists in

$$C([0, +\infty[; \widetilde{B}^s_{2,1}) \cap L^1([0, +\infty[; \widetilde{B}^{s+1}_{2,1}) \times L^\infty(0, \infty; \widetilde{B}^s_{2,1}) \cap L^1([0, +\infty[; \widetilde{B}^s_{2,1}).$$

The structure of this text will be the following.

- The second section is devoted to the proof of the local well-posedness. We use the very classical Friedrichs method.
- The third section consists of the proof of an a priori estimate for solutions of 2-D incompressible Navier–Stokes equations.
- The fourth section is the proof of a losing a priori estimate for solutions of transport equations. This means that we estimate the norm of the solutions of a transport equation in norms of Besov spaces whose index decreases in time; estimates of this type have been proved in [1] and in [8].
- In the fifth section, we study global well-posedness in the case of small data.

2. Local well-posedness and energy methods. We shall use the very classical Friedrichs method (also called the Galerkin method in the periodic case) which consists of an approximation of the system (VE) by a cut-off in the frequency space.

Let us define the operator J_n by

$$J_n a \overset{\text{def}}{=} \mathcal{F}^{-1}(\mathbf{1}_{B(0,n)}(\xi)\widehat{u}(\xi)),$$

where $\mathcal{F}$ denotes the Fourier transform in the space variables. Let us consider the approximate (VE_n) system

$$(VE_n) \begin{cases} \partial_t v_n - \nu J_n \Delta v_n &= J_n \mu_1 P \nabla \cdot \tau_n + J_n \mathcal{Q}(J_n v_n, J_n v_n), \\ \partial_t \tau_n + J_n(J_n v_n \cdot \nabla J_n \tau_n) + a J_n \tau_n &= \mu_2 D(v_n) + J_n Q(J_n \tau_n, J_n \nabla v_n), \\ \operatorname{div} v_n &= 0, \\ (v_n, \tau_n)_{|t=0} &= (J_n v_0, J_n \tau_0). \end{cases}$$

It is obvious that all the bilinear operators on the right are continuous on $L^2 \times L^2$. Then, the above system appears as a system of ordinary differential equations on L^2. Thus, the usual Cauchy–Lipschitz theorem implies the existence of a strictly positive maximal time T_n such that a unique solution exists which is continuous in time with value in L^2. However, as $J_n^2 = J_n$, we claim that $J_n(v_n, \tau_n)$ is also a solution, so uniqueness implies that $J_n(v_n, \tau_n) = (v_n, \tau_n)$. So (v_n, τ_n) is also a solution of the following system, still denoted by (VE_n):

$$(VE_n) \begin{cases} \partial_t v_n - \nu \Delta v_n &= \mu_1 P \nabla \cdot \tau_n + J_n \mathcal{Q}(v_n, v_n), \\ \partial_t \tau_n + J_n(v_n \cdot \nabla \tau_n) + a \tau_n &= \mu_2 D(v_n) + J_n Q(\tau_n, \nabla v_n), \\ \operatorname{div} v_n &= 0, \\ (v_n, \tau_n)_{|t=0} &= (J_n v_0, J_n \tau_0). \end{cases}$$

The system (VE_n) turns out to be an ordinary differential equation in L^2. So thanks to the Cauchy–Lipschitz theorem, a unique maximal solution exists on an interval $[0,T_n^\star[$ which is continuous in time with value in L^2. The main step consists of the proof of the following property:

For any real s strictly greater than $d/2$, a strictly positive constant c and a strictly positive time T exist so that, for any n, we have $T_n^\star \geq T$, and

$$\sup_{\substack{t\in[0,T]\\ n\in\mathbf{N}}} E_s((v_n,\tau_n),t)+c\nu\mu_2\int_0^T\|\nabla v_n(t)\|_{H^s}^2dt+c\mu_1 a\int_0^T\|\tau_n(t)\|_{H^s}^2dt\leq 2E_s((v,\tau),0), \tag{2.1}$$

where we defined

$$E_\sigma((v,\tau),t)\overset{\text{def}}{=}\mu_2\|v(t)\|_{H^\sigma}^2+\mu_1\|\tau(t)\|_{H^\sigma}^2.$$

Let us prove this estimate. As J_n are Fourier multipliers, they commute with constant coefficient differentiations; thus, applying the operator Δ_q to the system (VE_n), we obtain, by an energy estimate,

$$\begin{aligned}I_n(t)&\overset{\text{def}}{=}\frac{d}{dt}(\mu_2\|\Delta_q v_n(t)\|_{L^2}^2+\mu_1\|\Delta_q\tau_n\|_{L^2}^2)+2\mu_2\nu\|\nabla\Delta_q v_n(t)\|_{L^2}^2+2\mu_1 a\|\Delta_q\tau_n\|_{L^2}^2\\&=2\mu_2(\Delta_q J_n(v_n\cdot\nabla v_n)|\Delta_q v_n)_{L^2}-2\mu_1(\Delta_q J_n(v_n\cdot\nabla\tau_n)|\Delta_q\tau_n)_{L^2}\\&\qquad\qquad-2\mu_1(\Delta_q J_n Q(\tau u_n,\nabla v_n)|\Delta_q\tau_n)_{L^2}.\end{aligned}$$

Using that J_n is a real Fourier multiplier and that $J_n v_n=v_n$, we get

$$\begin{aligned}&\frac{d}{dt}(\mu_2\|\Delta_q v_n(t)\|_{L^2}^2+\mu_1\|\Delta_q\tau_n\|_{L^2}^2)+2\mu_2\nu\|\nabla\Delta_q v_n(t)\|_{L^2}^2+2\mu_1 a\|\Delta_q\tau_n\|_{L^2}^2\\&=2\mu_2(\Delta_q(v_n\cdot\nabla v_n)|\Delta_q v_n)_{L^2}-2\mu_1(\Delta_q(v_n\cdot\nabla\tau_n)|\Delta_q\tau_n)_{L^2}-2\mu_1(\Delta_q Q(\tau_n,\nabla v_n)|\Delta_q\tau_n)_{L^2}.\end{aligned}$$

Up to the end of the proof of the inequality (2.1), we shall drop the index n.

The classical tame estimates for the product in Sobolev spaces (see, for instance, Corollary 2.4.1 of [5]) imply that

$$\|Q(\tau,\nabla v)\|_{H^s}\leq C\Big(\|\tau(t)\|_{L^\infty}\|\nabla v(t)\|_{H^s}+\|\tau(t)\|_{H^s}\|\nabla v(t)\|_{L^\infty}\Big).$$

Thus, we infer that

$$\begin{aligned}&\Big|2\mu_1(\Delta_q Q(\tau,\nabla v)|\Delta_q\tau)_{L^2}\Big|\\&\leq C(1+|b|)\mu_1 2^{-qs}c_q(t)\Big(\|\tau(t)\|_{L^\infty}\|\nabla v(t)\|_{H^s}+\|\tau(t)\|_{H^s}\|\nabla v(t)\|_{L^\infty}\Big)\|\Delta_q\tau(t)\|_{L^2},\end{aligned} \tag{2.2}$$

where, as throughout this section, $c_q(t)$ denotes a positive series such that

$$\forall t\,,\ \sum_q c_q^2(t)=1.$$

Now we have to estimate terms of the type $(\Delta_q(v\cdot\nabla a)|\Delta_q a)_{L^2}$. To do so, we use Bony's decomposition and write that

$$v\cdot\nabla a=\sum_{j=1}^d\Big\{T_{v^j}\partial_j a+T_{\partial_j a}v^j+R(\partial_j a,v^j)\Big\}.$$

The classical results about paraproduct and remainder operator (see, for instance, Theorem 2.4.1. of [5]) imply that

$$\|T_{\partial_j a} v^j\|_{H^s} + \|R(\partial_j a, v^j)\|_{H^s} \leq C\|a\|_{L^\infty}\|\nabla v\|_{H^s}.$$

Thus we get

$$\left|2\mu_1\left(\Delta_q \sum_{j=1}^{d}(T_{\partial_j a}v^j + R(\partial_j a, v^j))|\Delta_q a\right)_{L^2}\right| \leq C\mu_1 2^{-qs}c_q(t)\|a(t)\|_{L^\infty}\|\nabla v(t)\|_{H^s}\|\Delta_q a(t)\|_{L^2}. \tag{2.3}$$

To estimate the last term, we start from a formula proved in [7]. Let us recall it:

$$\begin{aligned}\sum_{j=1}^{d}(\Delta_q T_{v^j}\partial_j a|\Delta_q a)_{L^2} = &\sum_{j,q'}([\Delta_q, S_{q'-1}v^j]\partial_j \Delta_{q'} a|\Delta_q a)_{L^2}\\ &+\frac{1}{2}\sum_{j,q',q''}\Big((S_{q''-1}v^j - S_{q'-1}v^j)\Delta_q\Delta_{q'}a|\partial_j\Delta_q\Delta_{q''}a\Big)_{L^2}.\end{aligned}$$

As we have, by definition of the operators Δ_q, that

$$[S_{q'-1}v^j, \Delta_q]b(x) = 2^{qd}\int (S_{q'-1}v^j(x) - S_{q'-1}v^j(y))h(2^q(x-y))b(y)dy,$$

we infer that

$$|[S_{q'-1}v^j, \Delta_q]b(x)| \leq C\|\nabla v\|_{L^\infty}2^{qd}\int |x-y||h(2^q(x-y)||b(y)|dy.$$

Thus we get that

$$\|[S_{q'-1}v^j, \Delta_q]b\|_{L^2} \leq C2^{-q}\|\nabla v\|_{L^\infty}\|b\|_{L^2}.$$

Moreover, thanks to Lemma 3.1, and using the fact that $|q'-q''| + |q'-q| \leq N_0$, we obtain that

$$\|(S_{q''-1}v^j - S_{q'-1}v^j)\Delta_q\Delta_{q'}a\|_{L^2} \leq C2^{-q}\|\nabla v\|_{L^\infty}\|\Delta_{q'}a\|_{L^2}.$$

Thus we get that

$$\left|2\mu_1\left(\Delta_q \sum_{j=1}^{d}T_{v^j}\partial_j a|\Delta_q a\right)_{L^2}\right| \leq C(1+|b|)\mu_1 2^{-qs}c_q(t)\|a(t)\|_{H^s}\|\nabla v(t)\|_{L^\infty}\|\Delta_q a(t)\|_{L^2} \tag{2.4}$$

Finally, applying estimates (2.2)–(2.4), we get

$$\begin{aligned}J_q(t) \overset{\text{def}}{=} &\frac{d}{dt}(\mu_2\|\Delta_q v(t)\|_{L^2}^2 + \mu_1\|\Delta_q\tau\|_{L^2}^2) + 2\mu_2\nu\|\nabla\Delta_q v_n(t)\|_{L^2}^2 + 2\mu_1 a\|\Delta_q\tau_n\|_{L^2}^2\\ \leq\ &Cc_q(t)2^{-qs}\Big(\|\nabla v(t)\|_{L^\infty}\mu_2\|v(t)\|_{H^s}\|\Delta_q v(t)\|_{L^2}\\ &\qquad + (1+|b|)\|\nabla v(t)\|_{L^\infty}\mu_1\|\tau(t)\|_{H^s}\|\Delta_q\tau(t)\|_{L^2}\\ &\qquad + \|\tau(t)\|_{L^\infty}\mu_1\|\nabla v(t)\|_{H^s}\|\Delta_q\tau(t)\|_{L^2}\Big).\end{aligned} \tag{2.5}$$

Then, summing in q gives

$$\frac{d}{dt}E_s(v,\tau)(t)+2\mu_2\nu\|\nabla v(t)\|_{H^s}^2+2\mu_1 a\|\tau_n\|_{H^s}^2\leq C(1+|b|)\|\nabla v(t)\|_{L^\infty}E_s(v,\tau)(t)$$
$$+\left(\frac{C\mu_1}{\mu_2\nu}\right)^{\frac{1}{2}}\|\tau(t)\|_{L^\infty}(\mu_2\nu)^{\frac{1}{2}}\|\nabla v(t)\|_{H^s}\mu_1^{\frac{1}{2}}\|\tau(t)\|_{H^s}.$$

So this implies that

$$\frac{d}{dt}E_s((v,\tau),t)+\mu_2\nu\|\nabla v(t)\|_{H^s}^2\leq C\left(\frac{\mu_1}{\mu_2\nu}\|\tau(t)\|_{L^\infty}^2+(1+|b|)\|\nabla v(t)\|_{L^\infty}\right)E_s((v,\tau),t).$$

The Gronwall lemma implies that, for any time t,

$$E_s((v,\tau),t)+\mu_2\nu\int_0^t\|\nabla v(t')\|_{H^s}^2dt'$$
$$\leq E_s((v,\tau),0)\exp C\int_0^t\left(\frac{\mu_1}{\mu_2\nu}\|\tau(t')\|_{L^\infty}^2+(1+|b|)\|\nabla v(t')\|_{L^\infty}\right)dt'.$$

Let us define T_n as

$$T_n\overset{\text{def}}{=}\sup\left\{t\,/\,\forall t'\leq t\,,\;E_s((v_n,\tau_n),t')+\mu_2\nu\int_0^{t'}\|\nabla v_n(t'')\|_{H^s}^2dt''\leq 2E_s((v,\tau),0)\right\}.$$

As $s>d/2$, the Sobolev embedding implies that, for any $t\leq T_n$,

$$\int_0^t\left(\frac{\mu_1}{\mu_2\nu}\|\tau_n(t')\|_{L^\infty}^2+\|\nabla v_n(t')\|_{L^\infty}\right)dt'$$
$$\leq C(1+|b|)\left(\frac{(tE_s((v,\tau),0)}{\mu_2\nu}\right)^{\frac{1}{2}}\left(1+\left(\frac{tE_s((v,\tau),0)}{\mu_2\nu}\right)^{\frac{1}{2}}\right)\cdot$$

Thus, it is easily inferred that if

$$T\leq\min\left\{T_n,\frac{\mu_2\nu}{C(1+|b|)E_s((v,\tau),0)}\right\},$$

then, for any $n\in N$ and any $t\in[0,T]$, we have

$$E_s((v_n,\tau_n),t)+\mu_2\nu\int_0^t\|\nabla v_n(t')\|_{H^s}^2dt'\leq 2E_s((v,\tau),0).$$

So this implies that, for any $n\in\mathbf{N}$,

$$t_n\geq\frac{\mu_2\nu}{CE_s((v,\tau),0)}\cdot$$

Standard compactness arguments imply the existence of a solution (v,τ) in $L_T^\infty(H^s)$ so that the vector field v belongs to $L^2([0,T];H^{s+1})$.

Now let us use the smoothing effect of the heat equation; let us consider a solution of (VE) which belongs to $L^\infty([0,T];H^s)$. Then it is obvious that

$$f\overset{\text{def}}{=}-v\cdot\nabla v+\nabla p+\nabla\cdot\tau\in L_T^\infty(H^{s-1}).$$

So the smoothing effect of the heat equation, as described for instance in [6], implies that if

$$\partial_t v - \nu \Delta v = f,$$

then, for any $p \in [1, \infty]$, we get

$$\|\Delta_q v\|_{L^p_T(L^2)} \leq \frac{C}{\nu} 2^{-q(s+1)} T^{\frac{1}{p}} \|f\|_{L^\infty_T(H^{s-1})} + \left(\frac{C}{2^{2q}\nu}\right)^{\frac{1}{p}} \|\Delta_q v(0)\|_{L^2}.$$

So taking $p = 1$ in the above estimate implies that, for any ε, the series $(\Delta_q v)_{q\in\mathbf{N}}$ is convergent in $L^1_T(H^{s+1-\varepsilon})$. Thus, for any strictly positive t_0, a strictly positive time t_1 exists so that $v(t_1, \cdot)$ belongs to $H^{s+1-\varepsilon}$. Let us apply the above estimate with $p = \infty$; we get that v belongs to $L^\infty([t_1, T]; H^{s+1-\varepsilon})$. Thanks to Sobolev embeddings, the fact that v is in $L^1_T(H^{s+1-\varepsilon})$ implies that v belongs to $L^1_T(Lip)$.

Let us prove the uniqueness. Let us consider two solutions (v_1, τ_1) and (v_2, τ_2) of (VE) in $L^\infty_{loc}([0, T^\star[; H^s)$. These solutions are such that v_j belongs to

$$L^2_{loc}([0, T^\star[; H^{s+1}) \cap L^\infty_{loc}(]0, T^\star[; H^{s+1-\varepsilon}).$$

Denoting by (w, θ) the difference between those two solutions, we have the following system:

$$\begin{cases} \dfrac{\partial w}{\partial t} - \nu\Delta w + v_2 \cdot \nabla w &= \mu_1 \nabla \cdot \theta - w \cdot \nabla v_1 - \nabla p, \\ \dfrac{\partial \theta}{\partial t} + v_1 \cdot \nabla \theta + a\theta &= \mu_2 D w - Q(\theta, \nabla v_1) - Q(\tau_2, \nabla w) - w \cdot \nabla \tau_2, \\ \operatorname{div} w &= 0. \end{cases}$$

By the L^2 energy estimate, we get

$$\frac{d}{dt}\big(\mu_2 \|w(t)\|^2_{L^2} + \mu_1 \|\theta(t)\|^2_{L^2}\big) + 2\mu_2 \nu \|\nabla w(t)\|^2_{L^2} + 2\mu_1 a \|\theta(t)\|^2_{L^2}$$
$$\leq 2\mu_1 \Big|(Q(\tau_2, \nabla w)|\theta)_{L^2} + (Q(\theta, \nabla v_1)|\theta)_{L^2}\Big| + 2\mu_1 \Big|(w \cdot \nabla \tau_2|\theta)_{L^2}\Big| + 2\mu_2 \Big|(w \cdot \nabla v_1|w)_{L^2}\Big|.$$

The following L^2 estimates are obvious:

$$\begin{aligned} |\mu_2(w \cdot \nabla v_1|w)_{L^2}| &\leq \mu_2 \|\nabla v_1||_{L^\infty} ||w||^2_{L^2}, \\ |\mu_1(Q(\theta, \nabla v_1)|\theta)_{L^2}| &\leq \mu_1(1 + |b|)\|\nabla v_1||_{L^\infty} \|\theta\|^2_{L^2}, \quad \text{and} \\ |\mu_1(Q(\tau_2, \nabla w)|\theta)_{L^2}| &\leq \mu_1(1 + |b|)\|\tau_2||_{L^\infty} \|\nabla w\|_{L^2} \|\theta\|_{L^2} \\ &\leq \frac{\nu\mu_2}{4} \|\nabla w\|^2_{L^2} + \frac{4\mu_1(1 + |b|)^2}{\mu_2 \nu} \|\tau_2\|^2_{L^\infty} \|\theta\|^2_{L^2}. \end{aligned}$$

To estimate the term $\mu_1(w \cdot \nabla \tau_2|\theta)_{L^2}$, we have to be a little more careful. Using the law of product in Sobolev spaces, we get, as s is strictly greater than $d/2$,

$$\|w \cdot \nabla \tau_2\|_{L^2} \leq \|\nabla w\|_{L^2} \|\tau_2\|_{H^s},$$

which yields

$$\begin{aligned} |\mu_1(w \cdot \nabla \tau_2|\theta)_{L^2}| &\leq \mu_1 \|\nabla w\|_{L^2} ||\tau_2||_{H^s} ||\theta||_{L^2} \\ &\leq \frac{\nu\mu_2}{4} ||\nabla w||^2_{L^2} + \frac{4\mu_1^2}{\mu_2 \nu}(1 + |b|)^2 ||\tau_2||^2_{H^s} ||\theta||^2_{L^2}. \end{aligned}$$

Plugging these estimates together, we infer

$$\begin{aligned}&\frac{d}{dt}\left(\mu_2\|w(t)\|_{L^2}^2+\mu_1\|\theta(t)\|_{L^2}^2\right)+\mu_2\nu\|\nabla w(t)\|_{L^2}^2+\mu_1 a\|\theta(t)\|_{L^2}^2\\ &\quad\leq C\left(\|\nabla v_1(t)\|_{L^\infty}+\frac{\mu_1}{\mu_2\nu}(1+|b|)^2\|\tau_2(t)\|_{H^s}^2\right)\left(\mu_2\|w(t)\|_{L^2}^2+\mu_1\|\theta(t)\|_{L^2}^2\right).\end{aligned}$$

So we get uniqueness by the Gronwall lemma.

3. Some a priori estimates for the 2-D Navier–Stokes system. Before stating Theorem 3.3 and Lemma 3.5, we recall some basic facts about Littlewood–Paley theory. We refer to [6] and [18] for the proof of the following results and for the multiplication law in Besov spaces.

Lemma 3.1.

$$\|\Delta_q u\|_{L^b}\leq 2^{d(\frac{1}{a}-\frac{1}{b})q}\|\Delta_q u\|_{L^a}\quad\text{for } b\geq a\geq 1,$$

$$\|e^{t\Delta}\Delta_q u\|_{L^b}\leq C2^{-ct2^{2q}}\|\Delta_q u\|_{L^b}.$$

Then the following corollary is obvious.

Corollary 3.1. *If $b\geq a\geq 1$, then we have the following continuous embeddings:*

$$B_{b,r}^{s-d\left(\frac{1}{a}-\frac{1}{b}\right)}\subset B_{a,r}^s.$$

Finally, we define the following space which will be used to control the system in dimension 2.

Definition 3.2. *Let p be in $[1,\infty]$ and r in $\mathbf{R}$; the space $\widetilde{L}_T^p(C^r)$ is the space of the distributions u such that*

$$\|u\|_{\widetilde{L}^p(0,T;C^r)}\overset{def}{=}\sup_q 2^{qr}\|\Delta_q u\|_{L_T^p(L^\infty)}<\infty.$$

Now let us state one of the two theorems about the 2-D Navier–Stokes system that we shall prove in this section.

Theorem 3.3. *Let v be the solution of the 2-D Navier–Stokes system with initial data in L^2 that belongs to $L_T^2(H^1)$ and an external force f in $L_T^1(C^{-1})\cap L_T^2(H^{-1})$; then, for any strictly positive ε, a T_0 in the interval $]0,T[$ exists such that*

$$\|\nabla v\|_{\widetilde{L}^1_{[T_0,T]}(C^0)}\leq\varepsilon.$$

The proof of this theorem will require two lemmas. The first one follows.

Lemma 3.4. *A constant C exists such that if v is the solution of the 2-D Navier–Stokes system with an initial data in L^2 that belongs to $L_T^2(H^1)$ and an external force f in $L_T^2(H^{-1})$, then*

$$\sum_q\|\Delta_q v\|_{L_T^\infty(L^2)}^2\leq\Big(1+\frac{C}{\nu}\|v_0\|_{L^2}^2\Big)\Big(\|v_0\|_{L^2}^2+\frac{C}{\nu}\|f\|_{L_T^2(H^{-1})}^2\Big).$$

To prove this lemma, we apply the operator Δ_q to the equation and, by energy estimate, we deduce that

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\Delta_q(t)\|_{L^2}^2+c\nu 2^{2q}\|\Delta_q(t)\|_{L^2}^2&\leq c\nu 2^{2q}\|\Delta_q v(t)\|_{L^2}^2\\ &\quad+\frac{C}{\nu}\Big(2^{-2q}\|\Delta_q f(t)\|_{L^2}^2+\|\Delta_q(v\otimes v)(t)\|_{L^2}^2\Big).\end{aligned}$$

So by integration and summation in q, we get

$$\sum_q \|\Delta_q v\|^2_{L^\infty_T(L^2)} \leq \|v_0\|^2_{L^2} + \frac{C}{\nu}(\|f\|^2_{L^2_T(H^{-1})} + \|v\|^4_{L^4_T(L^4)}).$$

Applying the classical inequality $\|a\|^4_{L^4} \leq C\|a\|^2_{L^2}\|a\|^2_{H^1}$, we get the result by the use of the standard energy estimate.

The second lemma is the following.

Lemma 3.5. *Let v be a solution of the Navier–Stokes system with initial data in L^2 and an external force f in $\widetilde{L}^1_T(C^{-1}) \cap L^2_T(H^{-1})$,*

$$(NS_\nu)\begin{cases} \dfrac{\partial v}{\partial t} + v\cdot\nabla v - \nu\Delta v &= -\nabla p + f,\\ \operatorname{div} v &= 0,\\ v|_{t=0} &= v_0. \end{cases}$$

Then we have the following a priori estimate:

$$\|v\|_{\widetilde{L}^1_T(C^1)} \leq CE_\nu(v_0,T) + \frac{C}{\nu}\|f\|_{\widetilde{L}^1_T(C^{-1})} + \frac{C}{\nu^2}\|\nabla v\|_{L^2_T(L^2)}\Big(\|v_0\|^2_{L^2} + \frac{2}{\nu}\|f\|^2_{L^2_T(H^{-1})}\Big),$$

where $E_\nu(v_0,T)$ is defined by

$$E_\nu(v_0,T) \overset{def}{=} \sup_q \|\Delta_q v_0\|_{L^2}\frac{1-e^{-c\nu T 2^{2q}}}{\nu}.$$

To prove this lemma, we first apply the operator Δ_q to the (NS_ν) system; this gives

$$\begin{aligned}\|\Delta_q v(t)\|_{L^\infty} \leq \|\Delta_q v_0\|_{L^\infty} e^{-\nu c 2^{2q} t} &+ \int_0^t e^{-\nu c 2^{2q}(t-t')}\|\Delta_q f(t')\|_{L^\infty} dt'\\ &+ \int_0^t e^{-\nu c 2^{2q}(t-t')}\|\Delta_q \mathcal{Q}(v(t'),v(t'))\|_{L^\infty} dt',\end{aligned}$$

with

$$\mathcal{Q}(v,v) = \sum_{i,j} A^{i,j}(D)(v^i v^j),$$

where $A^{i,j}(D)$ are homogeneous Fourier multipliers of degree 1. Using that

$$2^{-q}\|\Delta_q v_0\|_{L^\infty} \leq \|\Delta_q v_0\|_{L^2} \leq C\|v_0\|_{L^2}$$

thus yields, after integration in time, that

$$\|v\|_{\widetilde{L}^1_T(C^1)} \leq CE_\nu(v_0,T) + \frac{C}{\nu}\Big(\|\mathcal{Q}(v,v)\|_{\widetilde{L}^1_T(C^{-1})} + \|f\|_{\widetilde{L}^1_T(C^{-1})}\Big). \tag{3.1}$$

Denoting by $\widetilde{\varphi}^{i,j}(\xi) = \varphi(\xi)A^{i,j}(\xi) \in \mathcal{D}(\mathbf{R^2}\setminus\{\mathbf{0}\})$ we get

$$\|\Delta_q A^{i,j}(D)(v^i v^j)\|_{L^\infty} \leq C2^q\|\varphi^{i,j}(2^{-q}D)(v^i v^j)\|_{L^\infty}.$$

Then using the Bony's decomposition, we get

$$\begin{aligned}\varphi^{i,j}(2^{-q}D)(v^iv^j) &= \sum_{\substack{p'\geq q-2\\ |p-p'|\leq 2}} \varphi^{i,j}(2^{-q}D)(\Delta_p v^i \Delta_{p'} v^j)\\ &\quad+ \sum_{\substack{p'\geq q-2\\ p<p'-2}} \varphi^{i,j}(2^{-q}D)(\Delta_p v^i \Delta_{p'} v^j) + \sum_{\substack{p\geq q-2\\ p'<p-2}} \varphi^{i,j}(2^{-q}D)(\Delta_p v^i \Delta_{p'} v^j).\end{aligned}$$

For the first term, using the localization Lemma 3.1 and the fact that $|p-p'|\leq 2$, we have

$$\begin{aligned}\|\varphi^{i,j}(2^{-q}D)(\Delta_p v^i \Delta_{p'} v^j)\|_{L^\infty} &\leq 2^q\|\Delta_p v^i \Delta_{p'} v^j\|_{L^2} \leq C2^{q-p}2^{\frac{p}{2}}\|\Delta_p v^i\|_{L^2}2^{\frac{p'}{2}}\|\Delta_{p'}v^j\|_{L^\infty}\\ &\leq C2^{q-p'}\|v\|_{H^{\frac12}}2^{\frac{p'}{2}}\|\Delta_{p'}v^j\|_{L^\infty}.\end{aligned}$$

Hence summing up over p (a finite set for any fixed p'), integrating over $[0,T]$, and using the Hölder inequality, we get

$$\begin{aligned}\left\|\sum_{\substack{p'\geq q-2\\ |p-p'|\leq 2}} \varphi^{i,j}(2^{-q}D)(\Delta_p v^i \Delta_{p'} v^j)\right\|_{L^1_T(L^\infty)} &\leq \sum_{p'\geq q-2} C2^{q-p'}\|v\|_{L^4_T(H^{\frac12})}2^{\frac{p'}{2}}\|\Delta_{p'}v^j\|_{L^{\frac43}(L^\infty)}\\ &\leq C\|v\|_{L^4_T(H^{\frac12})}\|v\|_{\widetilde{L}^{\frac43}(C^{\frac12})},\end{aligned}\tag{3.2}$$

where we have used that $2^{\frac{p'}{2}}\|\Delta_{p'}v^j\|_{L^{\frac43}(L^\infty)} \leq \|v\|_{\widetilde{L}^{\frac43}(C^{\frac12})}$.

The second and the third terms are treated in the same way, we treat, for instance, the second one. We have

$$\begin{aligned}\|\varphi^{i,j}(2^{-q}D)(\Delta_p v^i \Delta_{p'} v^j)\|_{L^\infty} &\leq \|\Delta_p v^i\|_{L^\infty}\|\Delta_{p'}v^j\|_{L^\infty}\\ &\leq C2^p\|\Delta_p v^i\|_{L^2}\|\Delta_{p'}v^j\|_{L^\infty}\\ &\leq C2^{\frac{p-q}{2}}\|v\|_{H^{\frac12}}2^{\frac{p'}{2}}\|\Delta_{p'}v^j\|_{L^\infty}.\end{aligned}\tag{3.3}$$

Hence integrating over $[0,T]$, using the Hölder inequality, and noticing that the sum over p' can be restricted to the set $q+2\geq p'\geq q-2$ which is finite, we get

$$\left\|\sum_{\substack{p'\geq q-2\\ p\leq p'-2}} \varphi^{i,j}(2^{-q}D)(\Delta_p v^i \Delta_{p'} v^j)\right\|_{L^1_T(L^\infty)} \leq \sum_{p\leq q} C2^{\frac{p-q}{2}}\|v\|_{L^4_T(H^{\frac12})}\|v\|_{\widetilde{L}^{\frac43}_T(C^{\frac12})}.\tag{3.4}$$

Therefore, taking the supremum over q, we deduce that

$$\|\mathcal{Q}(v,v)\|_{\widetilde{L}^1_T(C^{-1})} \leq C\|v\|_{L^4_T(H^{\frac12})}\|v\|_{\widetilde{L}^{\frac43}_T(C^{\frac12})}.$$

Then by interpolation, we merely get that

$$\|v\|_{L^4_T(H^{\frac12})} \leq \|v\|^{1/2}_{L^2_T(H^1)}\|v\|^{1/2}_{L^\infty_T(L^2)}.$$

On the other hand, we have for all q

$$\begin{aligned}2^{\frac{q}{2}}\|\Delta_q v\|_{L^{\frac43}_T(L^\infty)} &\leq 2^{\frac{q}{2}}\|\Delta_q v\|^{1/2}_{L^1_T(L^\infty)}\|\Delta_q v\|^{1/2}_{L^2_T(L^\infty)}\\ &\leq 2^{\frac{q}{2}}\|\Delta_q v\|^{1/2}_{L^1_T(L^\infty)}\ 2^{\frac{q}{2}}\|\Delta_q v\|^{1/2}_{L^2_T(L^2)}\\ &\leq \|v\|^{1/2}_{\widetilde{L}^1_T(C^1)}\|v\|^{1/2}_{L^2_T(H^1)}.\end{aligned}$$

Thus, we infer that

$$\|\mathcal{Q}(v,v)\|_{\widetilde{L}^1_T(C^{-1})} \leq C\|v\|_{L^2_T(H^1)}\|v\|^{\frac{1}{2}}_{L^\infty_T(L^2)}\|v\|^{\frac{1}{2}}_{\widetilde{L}^1_T(C^1)}.$$

So plugging this estimate into (3.1), we obtain that

$$\|v\|_{\widetilde{L}^1_T(C^1)} \leq CE_\nu(v_0,T) + \frac{C}{\nu}\|f\|_{\widetilde{L}^1_T(C^{-1})} + C\|v\|_{L^2_T(H^1)}\|v\|^{\frac{1}{2}}_{L^\infty_T(L^2)}\|v\|^{\frac{1}{2}}_{\widetilde{L}^1_T(C^1)}.$$

But the energy estimate implies that

$$\|v\|_{L^\infty_T(L^2)} \leq \left(\|v_0\|^2_{L^2} + \frac{2}{\nu}\|f\|^2_{L^2_T(H^{-1})}\right)^{\frac{1}{2}}.$$

Thus, we infer that

$$\begin{aligned}\|v\|_{\widetilde{L}^1_T(C^1)} &\leq CE_\nu(v_0,T) + \frac{C}{\nu}\|f\|_{\widetilde{L}^1_T(C^{-1})} \\ &\qquad + \frac{C}{\nu}\|\nabla v\|_{L^2_T(L^2)}\left(\|v_0\|^2_{L^2} + \frac{2}{\nu}\|f\|^2_{L^2_T(H^{-1})}\right)^{\frac{1}{4}}\|v\|^{\frac{1}{2}}_{\widetilde{L}^1_T(C^1)} \\ &\leq CE_\nu(v_0,T) + \frac{C}{\nu}\|f\|_{\widetilde{L}^1_T(C^{-1})} \\ &\qquad + \frac{C}{\nu^2}\|\nabla v\|^2_{L^2_T(L^2)}\left(\|v_0\|^2_{L^2} + \frac{2}{\nu}\|f\|^2_{L^2_T(H^{-1})}\right)^{\frac{1}{2}} + \frac{1}{2}\|v\|_{\widetilde{L}^1_T(C^1)}.\end{aligned}$$

This concludes the proof of the lemma.

Now let us go the the proof of Theorem 3.3. First let us apply Lemma 3.5 between some T_0 in the interval $]0,T[$ and T. This gives

$$\begin{aligned}\|v\|_{\widetilde{L}^1_{[T_0,T]}(C^1)} \leq CE_\nu(v_{T_0},T-T_0) + \frac{C}{\nu}\|f\|_{L^1_{[T_0,T]}(C^{-1})} \\ + \frac{C}{\nu^2}\|\nabla v\|_{L^2_{[T_0,T]}(L^2)}\left(\|v_0\|^2_{L^2} + \frac{2}{\nu}\|f\|^2_{L^2_T(H^{-1})}\right).\end{aligned}$$

Lemma 3.4 implies in particular that, for any positive ε, an integer q_0 exists such that

$$\sup_{q\geq q_0}\|\Delta_q v\|_{L^\infty_T(L^2)} \leq \frac{\varepsilon\nu}{4C}\cdot$$

Then it turns out that

$$E_\nu(v_{T_0},T-T_0) \leq \frac{\varepsilon}{4C} + C\|v_0\|_{L^2}\nu 2^{2q_0}(T-T_0). \tag{3.5}$$

Now it is easy to choose T_0 such that, for any T' between T_0 and T, we get

$$\|f\|_{L^1_{[T',T]}(C^{-1})} \leq \frac{\varepsilon\nu}{4C} \quad \text{and} \quad \|\nabla v\|_{L^2_{[T',T]}(L^2)} \leq \frac{\varepsilon\nu^2}{4C}\left(\|v_0\|^2_{L^2} + \frac{2}{\nu}\|f\|^2_{L^2_T(H^{-1})}\right)^{-1}.$$

4. A losing a priori estimate. The core of this section is the proof of a losing estimate for transport equation in the spirit of [1]. After this proof, we shall apply this estimate in order to prove Theorem 1.2.

THEOREM 4.1. *Let σ and β be two elements of* $]0,1[$. *A constant C exists that satisfies the following properties.*

Let T and λ be two positive numbers and v a smooth divergence-free vector field so that

$$\sigma - \lambda \|\nabla v\|_{\widetilde{L}^1_T(C^0)} \geq \beta. \tag{4.1}$$

Consider two smooth functions f and g so that f is the solution of

$$(T)\left\{\begin{array}{rcl} \partial_t f + v\cdot\nabla f + Q(\nabla v, f) &=& g,\\ f_{|t=0} &=& f_0.\end{array}\right.$$

Then we have, if $\lambda \geq 2C$,

$$M^\sigma_\lambda(f) \leq 2\|f_0\|_{B^\sigma_{p,\infty}} + \frac{2C}{\lambda} M^{\sigma+1}_\lambda(v) + TM^\sigma_\lambda(g), \tag{4.2}$$

where

$$(4.3) \quad M^\sigma_\lambda(c) \overset{def}{=} \sup_{\substack{t\in[0,T]\\ q}} 2^{q\sigma-\Phi_{q,\lambda}(t)} \|\Delta_q c(t)\|_{L^p} \quad \textit{with}$$

$$(4.4)\, \Phi_{q,\lambda}(t,t') \overset{def}{=} \lambda \int_{t'}^t \|S_q\nabla v(t'')\|_{L^\infty} dt'' + \lambda \int_{t'}^t \|f(t'')\|_{L^\infty} dt''\,,\ \Phi_{q,\lambda}(t) = \Phi_{q,\lambda}(t,0).$$

To prove this theorem, we transform the transport equation (T) along the flow of v, in the following equation (T_q) on $f_q \overset{\text{def}}{=} \Delta_q f$, which is a transport equation along the flow of $S_q v$.

$$(T_q)\left\{\begin{array}{rcl} \partial_t f_q + S_q v\cdot\nabla f_q &=& \Delta_q g - R_q(v,f),\\ f_{q|t=0} &=& \Delta_q f_0.\end{array}\right.$$

Let us admit for a while the following estimate:

$$2^{q\sigma-\Phi_{q,\lambda}(t)} \|R_q(v(t),f(t))\|_{L^p} \leq Ce^{C\lambda\|\nabla v\|_{\widetilde{L}^1_T(C^0)}}$$

$$(4.5)\times\left(\|f(t)\|_{L^\infty} M^{\sigma+1}_\lambda(v) + \left(\|S_q\nabla v(t)\|_{L^\infty} + \sum_{|q'-q|\leq N} \|\Delta_{q'}\nabla v(t)\|_{L^\infty}\right) M^\sigma_\lambda(f)\right).$$

Let us denote by ψ_q the flow of the vector field $S_q v$. The equation (T_q) may be rewritten as

$$(\widetilde{T}_q) \qquad \frac{d}{dt} f_q(t,\psi_q(t,x)) = \Delta_q g(t,\psi_q(t,x)) - R_q(v(t),f(t))(\psi_q(t,x)).$$

As the vector field v, and of course also the vector field $S_q v$ is divergence-free, we get, after time integration in $(\widetilde{T}_q)$, that

$$\|f_q(t)\|_{L^p} \leq \|f_q(0)\|_{L^p} + \int_0^t \|\Delta_q g(t')\|_{L^p} dt' + \int_0^t \|R_q(v(t'),f(t'))\|_{L^p} dt'.$$

After a multiplication by $2^{q\sigma-\Phi_{q,\lambda}(t)}$, we get

$$\begin{aligned}2^{q\sigma-\Phi_{q,\lambda}(t)}\|f_q(t)\|_{L^p} \leq 2^{q\sigma}\|\Delta_q f_0\|_{L^p} + \int_0^t 2^{-\Phi_{q,\lambda}(t,t')}2^{q\sigma-\Phi_{q,\lambda}(t')}\|\Delta_q g(t')\|_{L^p}dt' \\ + \int_0^t 2^{-\Phi_{q,\lambda}(t,t')}2^{q\sigma-\Phi_{q,\lambda}(t')}\|R_q(v(t'),f(t'))\|_{L^p}dt'.\end{aligned}$$

Then, using the inequality (4.5), we get

$$\begin{aligned}M^\sigma_\lambda(f) \leq \|f_0\|_{B^\sigma_{p,\infty}} + TM^\sigma_\lambda(g) + e^{C\lambda\|\nabla v\|_{\widetilde{L}^1_T(C^0)}}M^\sigma_\lambda(f)\sup_{\substack{t\in[0,T]\\ q}}\int_0^t 2^{-\Phi_{q,\lambda}(t,t')} \\ \times\left(\|f(t')\|_{L^\infty}M^{\sigma+1}_\lambda(v) + M^\sigma_\lambda(f)\left(\|S_q\nabla v(t')\|_{L^\infty} + \sum_{|q'-q|\leq N}\|\Delta_{q'}\nabla v(t')\|_{L^\infty}\right)\right)dt'.\end{aligned}$$

As $\lambda\|\nabla v\|_{\widetilde{L}^1_T(C^0)}$ is smaller than $(\sigma-\beta)$, we have

$$e^{C\lambda\|\nabla v\|_{\widetilde{L}^1_T(C^0)}} \leq e^{C(\sigma-\beta)}.$$

Moreover, by definition of $\Phi_{q,\lambda}(t,t')$, it is obvious that

$$\int_0^t 2^{-\Phi_{q,\lambda}(t,t')}\|f(t')\|_{L^\infty}dt' \leq \frac{1}{\lambda\log 2} \quad\text{and}\quad \int_0^t 2^{-\Phi_{q,\lambda}(t,t')}\|S_q\nabla v(t')\|_{L^\infty}dt' \leq \frac{1}{\lambda\log 2}\cdot$$

Then we obtain that

$$\begin{aligned}M^\sigma_\lambda(f) &\leq \|f_0\|_{B^\sigma_{p,\infty}} + \frac{C}{\lambda}M^{\sigma+1}_\lambda(v) + C\|\nabla v\|_{\widetilde{L}^1_T(C^0)}M^\sigma_\lambda(f) + TM^\sigma_\lambda(g) + \frac{C}{\lambda}M^\sigma_\lambda(f) \\ &\leq \|f_0\|_{B^\sigma_{p,\infty}} + \frac{C}{\lambda}M^{\sigma+1}_\lambda(v) + TM^\sigma_\lambda(g) + \frac{C}{\lambda}M^\sigma_\lambda(f).\end{aligned}$$

This proves the theorem of course if we prove the estimate (4.5). First of all, let us decompose the operator R_q. We have

$$\begin{aligned}R_q(v,f) &= \sum_{\ell=1}^{8}R^\ell_q(v,f) \quad\text{with} \\ R^1_q(v,f) &= \sum_{j=1}^{d}\Delta_q(T_{\partial_j f}v^j), \\ R^2_q(v,f) &= \sum_{j=1}^{d}[\Delta_q, T_{v^j}\partial_j]f, \\ R^3_q(v,f) &= \sum_{j=1}^{d}T_{(v^j-S_qv^j)}\partial_j\Delta_q f, \\ R^4_q(v,f) &= \sum_{j=1}^{d}-T_{\partial_j\Delta_q f}S_qv^j, \\ R^5_q(v,f) &= \sum_{j=1}^{d}\Delta_q\partial_j R(v^j,f) - R(S_qv^j,\Delta_q\partial_j f),\end{aligned}$$

$$R_q^6(v,f)=\sum_{j=1}^d \Delta_q Q(T_{\nabla v},f),$$

$$R_q^7(v,f)=\sum_{j=1}^d \Delta_q Q(T_f,\nabla v),\quad \text{and}$$

$$R_q^8(v,f)=\sum_{j=1}^d \Delta_q Q(R(\nabla v,f)).$$

If f is a solution of (T), then

$$\partial_t f_q+\Delta_q(v\cdot\nabla f)+\Delta_q Q(\nabla v,f)=\Delta_q g.$$

Now, let us use Bony's decomposition of the products $v^j\partial_j f$. We thus get

$$(4.6)\ v^j\partial_j f=T_{v^j}\partial_j f+T_{\partial_j f}v^j+R(v^j,\partial_j f)\quad\text{and}\quad \Delta_q Q(\nabla v,f)=\sum_{\ell=6}^{8}R_q^\ell(v,f).$$

Then we have the following equalities:

$$\begin{aligned}\Delta_q(v\cdot\nabla f)&=R_q^1(v,f)+\sum_{j=1}^d \Delta_q T_{v^j}\partial_j f+\Delta_q R(v^j,\partial_j f)\\&=\sum_{\ell=1}^2 R_q^\ell(v,f)+\sum_{j=1}^d T_{v^j}\partial_j\Delta_q f+\Delta_q R(v^j,\partial_j f),\\&=\sum_{\ell=1}^3 R_q^\ell(v,f)+\sum_{j=1}^d T_{S_q v^j}\partial_j\Delta_q f+\Delta_q R(v^j,\partial_j f).\end{aligned}$$

Then, using the definition of the paraproduct and the fact that the vector field v is divergence-free, we infer that

$$\Delta_q(v\cdot\nabla f)=\sum_{\ell=1}^5 R_q^\ell(v,f)+S_q v\cdot\nabla\Delta_q f.$$

Now let us prove that each term $R_q^\ell(v,f)$ can be estimated with the right term of inequality (4.5).

Let us begin with $R_q^1(v,f)$. By definition of the paraproduct, we have

$$R_q^1(v,f)=\sum_{j=1}^d\sum_{q'}\Delta_q(S_{q'-1}\partial_j f\Delta_{q'}v^j).$$

As $|q-q'|>N$, the above term is then equal to 0, and we deduce that

$$\|R_q^1(v(t),f(t))\|_{L^p}\le C\sum_{|q-q'|\le N}\|S_{q'-1}\nabla f\|_{L^\infty}\|\Delta_{q'}v(t)\|_{L^p}.$$

Using the fact that, if $|q-q'|\le N$, then $\|S_{q'-1}\nabla f\|_{L^\infty}\le C2^q\|f(t)\|_{L^\infty}$, and we infer that

$$\|R_q^1(v(t),f(t))\|_{L^p}\le C2^q\|f(t)\|_{L^\infty}\sum_{|q-q'|\le N}\|\Delta_{q'}v(t)\|_{L^p}.$$

So we claim that

$$\begin{aligned}&2^{q\sigma-\Phi_{q,\lambda}(t)}\|R_q^1(v(t),f(t))\|_{L^p}\\&\quad\leq C\|f(t)\|_{L^\infty}M_\lambda^{\sigma+1}(v)\sum_{|q-q'|\leq N}2^{-\lambda\int_0^t\|S_q\nabla v(t')\|_{L^\infty}dt'+\lambda\int_0^t\|S_{q'}\nabla v(t')\|_{L^\infty}dt'}.\end{aligned}$$

However, it is obvious that

$$\int_0^t\|S_{q'}\nabla v(t')\|_{L^\infty}dt'-\int_0^t\|S_q\nabla v(t')\|_{L^\infty}dt'\leq\int_0^t\|(S_{q'}-S_q)\nabla v(t')\|_{L^\infty}dt'.$$

Using the fact that $|q-q'|\leq N_0$ and that

$$\|\Delta_q u\|_{L^p}\leq C2^{-q}\|\nabla\Delta_q u\|_{L^p},$$

we get

$$\int_0^t\|S_{q'}\nabla v(t')\|_{L^\infty}dt'-\int_0^t\|S_q\nabla v(t')\|_{L^\infty}dt'\leq C\|\nabla v\|_{\widetilde{L}^1_T(C^0)}. \tag{4.7}$$

Thus it turns out that

$$2^{q\sigma-\Phi_{q,\lambda}(t)}\|R_q^1(v(t),f(t))\|_{L^p}\leq C\|f(t)\|_{L^\infty}2^{C\lambda\|\nabla v\|_{\widetilde{L}^1_T(C^0)}}M_\lambda^{\sigma+1}(v). \tag{4.8}$$

Now let us look at $R_q^2(v,f)$. By definition of the paraproduct, we have

$$\begin{aligned}R_q^2(v,f)&=-\sum_{j=1}^d\sum_{q'}[S_{q'-1}v^j\partial_j\Delta_{q'},\Delta_q]f\\&=-\sum_{j=1}^d\sum_{q'}[S_{q'-1}v^j,\Delta_q]\partial_j\Delta_{q'}f.\end{aligned}$$

The terms of the above sum are equal to 0 except if $|q-q'|\leq N$. Moreover, by definition of the operators Δ_q, we have

$$[S_{q'-1}v^j,\Delta_q]\partial_j\Delta_{q'}f(x)=2^{qd}\int_{\mathbf{R^d}}h(2^q(x-y))(S_{q'-1}v^j(x)-S_{q'-1}v^j(y))\partial_j\Delta_{q'}f(y)dy.$$

So we infer that

$$|[S_{q'-1}v^j,\Delta_q]\partial_j\Delta_{q'}f(x)|\leq2^{-q}\|\nabla S_{q'-1}v\|_{L^\infty}2^{qd}\Big(\Big(2^q|\cdot|\times|h(2^q\cdot)|\Big)\star|\partial_j\Delta_{q'}f|\Big)(x).$$

Then we have, using inequality (4.7),

$$\begin{aligned}&2^{q\sigma-\Phi_{q,\lambda}(t)}\|[S_{q'-1}v^j,\Delta_q]\partial_j\Delta_{q'}f\|_{L^p}\\&\quad\leq CM_\lambda^\sigma(f)\sum_{|q-q'|\leq N}2^{C\lambda\|v\|_{\widetilde{L}^1_T(C^1)}}(\|\nabla(S_{q'-1}-S_q)v(t)\|_{L^\infty}+\|S_qv(t)\|_{L^\infty}).\end{aligned}$$

We thus get

$$\begin{aligned}&2^{q\sigma-\Phi_{q,\lambda}(t)}\|R_q^2(v(t),f(t))\|_{L^p}\leq CM_\lambda^\sigma(f)2^{C\lambda\|v\|_{\widetilde{L}^1(C^1)}}\\&\qquad\times\left(\|S_q\nabla v(t)\|_{L^\infty}+\sum_{|q-q'|\leq N}\|\nabla(S_{q'-1}-S_q)v(t)\|_{L^\infty}\right).\end{aligned} \tag{4.9}$$

The estimate about $R^3_q(v,f)$ is very easy to prove. By definition of the paraproduct, we have

$$R^3_q(v,f) = \sum_{q'} (S_{q'-1}v^j - S_q v^j)\Delta_{q'}\Delta_q f,$$

so we get

$$\|R^3_q(v,f)\|_{L^p} \leq C \sum_{q' \geq q} 2^{q-q'} \|\Delta_{q'}\nabla v(t)\|_{L^\infty} \|D_q f(t)\|_{L^p}.$$

So by definition of $M^\sigma_\lambda(f)$ it is obvious that

$$2^{q\sigma-\Phi_{q,\lambda}(t)}\|R^3_q(v(t),f(t))\|_{L^p} \leq CM^\sigma_\lambda(f) \sum_{q'\geq q} 2^{q-q'}\|\Delta_{q'}v(t)\|_{L^\infty}. \tag{4.10}$$

Now let us estimate $R^4_q(v,f)$. By definition of the paraproduct, we have

$$R^4_q(v,f) = \sum_{j=1}^{d}\sum_{q'} S_{q'-1}\Delta_q\partial_j f \Delta_{q'} S_q v^j.$$

It is obvious by definition of the operators S_q and Δ_q that if $q' \leq q$, then $S_{q'-1}\Delta_q = 0$ and if $q' \geq q+1$, then $\Delta_{q'}S_q = 0$. So

$$R^4_q(v,f) = \sum_{j=1}^{d} S_{q-1}\Delta_q\partial_j f\Delta_q S_q v^j.$$

It turns out that

$$2^{q\sigma-\Phi_{q,\lambda}(t)}\|R^4_q(v(t),f(t))\|_{L^p} \leq CM^\sigma_\lambda(f)\|S_q\nabla v(t)\|_{L^\infty}. \tag{4.11}$$

The estimate of $R^5_q(v,f)$ is a little bit more delicate. We have, using the fact that v is divergence-free,

$$\begin{aligned}
R^5_q(v,f) &= \sum_{\ell=1}^{2} R^{5,\ell}_q(v,f) \quad \text{with}\\
R^{5,1}_q(v,f) &= \sum_{j=1}^{d} \partial_j\Delta_q R((\mathrm{Id}-S_q)v^j, f) \quad \text{and}\\
R^{5,2}_q(v,f) &= \sum_{j=1}^{d} \Delta_q R(S_q v^j, \partial_j f) - R(S_q v^j, \partial_j\Delta_q f).
\end{aligned}$$

The estimate of $R^{5,1}_q(v,f)$ is analogous to the one of $R^5_q(v,f)$. We get

$$2^{qs-\Phi_{q,\lambda}(t)}\|R^{5,1}_q(v(t),f(t))\|_{L^p} \leq Ce^{C\lambda\|\nabla v\|_{\widetilde{L}^1_T(C^0)}} M^s_\lambda(f)\sum_{q'\geq q} 2^{q-q'}2^{q'}\|\Delta_{q'}v(t)\|_{L^\infty}. \tag{4.12}$$

By definition of the remainder operators, it turns out that

$$R^{5,2}_q(v,f) = \sum_{\substack{|q'-q''|\leq 1\\ q'\geq q-N}} [\Delta_q, \Delta_{q'}S_q(v^j)]\Delta_{q''}\partial_j f.$$

Always along the same lines, we get

$$2^{qs-\Phi_{q,\lambda}(t)}\|R_q^{5,2}(v(t),f(t))\|_{L^p} \leq Ce^{C\lambda\|\nabla v\|_{\widetilde{L}^1_T(C^0)}} M^s_\lambda(f)\sum_{q'\geq q} 2^{q-q'}2^{q'}\|\Delta_{q'}v(t)\|_{L^\infty}. \tag{4.13}$$

The term $R_q^6(v,f)$ is estimated exactly as the term $R_q^1(v,f)$, the term $R_q^7(v,f)$ exactly as the term $R_q^3(v,f)$, and the term $R_q^8(v,f)$ exactly as the term $R_q^5(v,f)$. So putting together estimates (4.8)–(4.13), we get the estimate (4.5) and thus Theorem 4.1.

Now we return to the proof of Theorem 1.2. We assume that we have a solution given by Theorem 1.1 on an interval $[0,T[$. Let us assume that

$$T<\infty \quad \text{and} \quad \int_0^T (\|\tau(t,\cdot)\|_{L^\infty}+b\|\tau(t)\|^2_{L^2})dt<\infty.$$

We want to prove that we can prolong the solution.

Theorem 1.1 says that, for any T_0 in $]0,T[$, the solution (v,τ) of (VE) belongs to the space $L^\infty_{loc}([T_0,T[;H^{s+1}\times H^s)$. Sobolev-type embeddings of Corollary 3.1 imply that

$$(v,\tau)\in L^\infty_{loc}\Big([T_0,T[;\widetilde{B}^{s+1-2\left(\frac12-\frac1p\right)}_{p,\infty}\times \widetilde{B}^{s-2\left(\frac12-\frac1p\right)}_{p,\infty}\Big).$$

Choosing $p=\infty$ in the above assertion implies that $(v,\tau)\in L^\infty_{loc}(\widetilde{C}^s\times\widetilde{C}^{s-1})$. As s is greater than 1, the tensor τ belongs to $L^2([T_0,T];L^2)\cap L^1([T_0,T];C^0)$. So we can apply Theorem 3.3. We thus choose T_0 such that, with the notations of Theorem 4.1, we have

$$\|\nabla v\|_{\widetilde{L}^1_{[T_0,T]}(C^0)}\leq \frac{s-1-\beta}{2\lambda}.$$

The losing estimate of Theorem 4.1 applied with $\sigma=s-1$ and between T_0 and T says exactly that the tensor τ satisfies

$$M^{s-1}_\lambda(\tau)\leq 2\|\tau_0\|_{C^{s-1}}+\left(\frac{C}{\lambda}+T-T_0\right)M^s_\lambda(v). \tag{4.14}$$

Now we have to estimate ∇v. The 2-D Navier–Stokes equation can be written as

$$\partial_t v-\nu\Delta v=P(v\cdot\nabla v)+PD\tau,$$

wherein P denotes the Leray projector on the divergence-free vector field. Along the exact same lines as in the proof of Theorem 4.1, we have

$$\begin{aligned}&2^{qs-\Phi_{q,\lambda}(t)}\|P(v\cdot\nabla v)-P(S_qv\cdot\nabla\Delta_qv)\|_{L^\infty}\\ &\qquad\leq CM^s_\lambda(v)\left(\|S_q\nabla v(t)\|_{L^\infty}+\sum_{q'\geq q}2^{q-q'}\|\nabla\Delta_{q'}v(t)\|_{L^\infty}\right).\end{aligned}\tag{4.15}$$

Moreover, it is obvious that

$$2^{q(s-\frac32)-\Phi_{q,\lambda}(t)}\|P(S_qv\cdot\nabla\Delta_qv)\|_{L^\infty}\leq C\|v(t)\|_{H^{\frac12}}M^s_\lambda(v).$$

So it turns out that

$$
\begin{aligned}
&2^{qs-\Phi_{q,\lambda}(t)}\|\Delta_q P(v\cdot\nabla v)\|_{L^\infty}\\
&\le CM^s_\lambda(v)\left(\|S_q\nabla v(t)\|_{L^\infty}+\sum_{q'\ge q}2^{(q-q'}\|\nabla v(t)\|_{L^\infty}+2^{\frac{q}{2}}\|v(t)\|_{H^{\frac12}}\right).
\end{aligned}
\tag{4.16}
$$

Using well-known estimates on the heat equation (see, for instance, [6]) and inequalities (4.14) and (4.16), we get that

$$M^s_\lambda(v)\le\|v_0\|_{C^s}+2\|\tau_0\|_{C^{s-1}}+\left(\frac{C}{\lambda}+T-T_0+2^{\frac{q}{2}}F_q(T_0,T)\right)M^s_\lambda(v)$$

with

$$F_q(T_0,T)\stackrel{\text{def}}{=}\sup_{t\in[T_0,T]}\int_{T_0}^t e^{c\nu 2^{2q}(t-t')}\|v(t')\|_{H^{\frac12}}\,dt'.$$

Hölder inequality implies immediately that

$$F_q(T_0,T)\le\frac{C}{\nu^{\frac34}}2^{-\frac{q}{2}}\|v\|_{L^4_{T_0,T]}(H^{\frac12})}.$$

We thus infer that

$$M^s_\lambda(v)\le\|v_0\|_{C^s}+2\|\tau_0\|_{C^{s-1}}+\left(\frac{C}{\lambda}+T-T_0+\frac{C}{\nu^{\frac34}}\|v\|_{L^4_{T_0,T]}(H^{\frac12})}\right)M^s_\lambda(v).$$

Now it is enough to choose T_0 such that the quantity

$$\frac{C}{\lambda}+T-T_0+\frac{C}{\nu^{\frac34}}\|v\|_{L^4_{[T_0,T]}(H^{\frac12})}$$

is small enough. Then as s is greater than 1, the solution (v,τ) of the system (VE) is such that $(\nabla v,\tau)$ belongs to $L^\infty([T_0,T^\star]\times\mathbf{R^2})$; this concludes the proof of Theorem 1.2.

5. Local and global existence for initial data in Besov spaces. The proof of Theorems 1.6 and 1.7 is based on the following lemma.

LEMMA 5.1. *Let s be in the interval $]-s_c,s_c+1]$. A constant C exists which satisfies the following properties.*

Let us consider any divergence-free vector field v in $L^1_T(B^{s_c+1}_{p,1})$ and any solution (w,τ) of the following linear system:

$$
(VEL)\left\{
\begin{array}{rcl}
\partial_t w-\nu\Delta w &=& P\mu_1\nabla\tau+f,\\
\partial_t\tau+v\cdot\nabla\tau+Q(\tau,\nabla v)+a\tau &=& \mu_2 Dw+g,\\
(w,\tau)_{|t=0} &=& (w_0,\tau_0)
\end{array}
\right.
$$

with $(w_0,\tau_0)\in B^{s-1}_{p,1}\times B^s_{p,1}$ and $(f,g)\in L^1_T(B^{s-1}_{p,1})\times L^1_T(B^s_{p,1})$.

Let us define

$$T_{VE}\stackrel{def}{=}+\infty\ \ \textit{if}\ \ \mu_1\mu_2\le c\nu a\qquad\textit{and}\qquad T_{VE}\stackrel{def}{=}-\frac1a\log\left(1-\frac{c\nu a}{2\mu_1\mu_2}\right)\ \ \textit{if}\ \ \mu_1\mu_2\ge c\nu a.\tag{5.1}$$

Then, if $T \leq T_{VE}$, we get for some λ big enough

$$\begin{aligned} \|w_\lambda\|_{L^1_T(B^{s+1}_{p,1})} \leq \frac{C}{\nu} \Bigg(& \sum_q 2^{q(s-1)} \|\Delta_q w_0\|_{L^p} (1 - e^{C^{-1}\nu 2^{2q}T}) + \|f_\lambda\|_{L^1_T(B^{s-1}_{p,1})} \\ & + \mu_1 \frac{1 - e^{-aT}}{a} \Big(\|\tau_0\|_{B^s_{p,1}} + \|g_\lambda\|_{L^1_T(B^s_{p,1})} \Big) \Bigg), \end{aligned} \tag{5.2}$$

$$\begin{aligned} \|w_\lambda\|_{L^\infty_T(B^{s-1}_{p,1})} \leq C \Bigg(& \|w_0\|_{B^{s-1}_{p,1}} + \|f_\lambda\|_{L^1_T(B^{s-1}_{p,1})} \\ & + \mu_1 \frac{1 - e^{-aT}}{a} \Big(\|\tau_0\|_{B^s_{p,1}} + \|g_\lambda\|_{L^1_T(B^s_{p,1})} \Big) \Bigg), \end{aligned} \tag{5.3}$$

$$\begin{aligned} \|\tau_\lambda\|_{L^\infty_T(B^s_{p,1})} \leq & \, C \left(1 + \frac{\mu_1 \mu_2}{\nu a} (1 - e^{-aT}) \right) \left(\|\tau_0\|_{B^s_{p,1}} + \|g_\lambda\|_{L^1_T(B^s_{p,1})} \right) \\ & + \frac{C \mu_2}{\nu} \left(\sum_q 2^{q(s-1)} \|\Delta_q w_0\|_{L^p} (1 - e^{C^{-1}\nu 2^{2q}T}) + \|f_\lambda\|_{L^1_T(B^{s-1}_{p,1})} \right) \end{aligned} \tag{5.4}$$

with

$$a_\lambda(t) \overset{def}{=} a(t) \exp \left(-\lambda \int_0^t \|\nabla v(t')\|_{B^{s_c}_{p,1}} dt' \right).$$

Remark. The condition (1.8) $\mu_1 \mu_2 \leq c \nu a$ means that the coupling effect between the two equations is less important than the viscosity effect on the time interval $[0, T]$. Let us also note that, in any case, spatially if $a = 0$, we may take

$$T_{VE} = \frac{\nu}{2C\mu_1\mu_2} \cdot$$

To prove the above lemma, we start with an estimate on τ (inequality (5.6) below) and then plug it into a standard estimate on w.

We apply the operator Δ_q on the transport equation on τ; we thus get

$$\begin{aligned} \partial_t \Delta_q \tau + v \cdot \nabla \Delta_q \tau + a \Delta_q \tau \ &= \ \mu_2 D \Delta_q w + \Delta_q g + R_q(v, \tau) \quad \text{with} \\ R_q(v, \tau) &\overset{\text{déf}}{=} [v \cdot \nabla, \Delta_q] \tau - \Delta_q Q(\tau, \nabla v). \end{aligned}$$

Let us admit for a while the following estimate:

$$\|R_q(v(t), \tau(t))\|_{L^p} \leq C 2^{-qs} c_q(t) \|\nabla v(t)\|_{B^{s_c+1}_{p,1}} \|\tau(t)\|_{B^s_{p,1}}, \tag{5.5}$$

where, as all along this section, $c_q(t)$ denote a positive series whose sum over q is 1. As the vector field v is divergence-free, we get, using integration along the characteristics that

$$\begin{aligned} e^{at} \|\Delta_q \tau(t)\|_{L^p} \leq \|\Delta_q \tau_0\|_{L^p} + C 2^{-qs} & \int_0^t c_q(t') \|\nabla v(t')\|_{B^{s_c+1}_{p,1}} e^{at'} \|\tau(t')\|_{B^s_{p,1}} dt' \\ & + C \mu_2 2^q \int_0^t e^{at'} \|\Delta_q w(t')\|_{L^p} dt' + C \int_0^t e^{at'} \|\Delta_q g(t')\|_{L^p} dt'. \end{aligned}$$

Then, using the multiplication by $2^{qs}\exp\Big(-\lambda\int_0^t\|\nabla v(t')\|_{B^{s_c}_{p,1}}dt'\Big)$, we get

$$\begin{aligned}e^{at}\|\Delta_q\tau_\lambda(t)\|_{L^p}\leq{}&\|\Delta_q\tau_0\|_{L^p}\\&+C\int_0^t c_q(t')e^{-\lambda\int_{t'}^t\|\nabla v(t'')\|_{B^{s_c}_{p,1}}dt''}\|\nabla v(t')\|_{B^{s_c}_{p,1}}e^{at'}\|\tau_\lambda(t')\|_{B^s_{p,1}}dt'\\&+C\mu_2 2^q\int_0^t e^{at'}2^{q(s+1)}\|\Delta_q w_\lambda(t')\|_{L^p}dt'+C\int_0^t e^{at'}2^{qs}\|\Delta_q g_\lambda(t')\|_{L^p}dt'.\end{aligned}$$

Taking the sum over q, we get

$$\begin{aligned}e^{at}\|\tau_\lambda(t)\|_{B^s_{p,1}}\leq\|\tau_0\|_{B^s_{p,1}}&+C\int_0^t e^{-\lambda\int_{t'}^t\|\nabla v(t'')\|_{B^{s_c}_{p,1}}dt''}\|\nabla v(t')\|_{B^{s_c}_{p,1}}e^{at'}\|\tau_\lambda(t')\|_{B^s_{p,1}}dt'\\&+C\mu_2\int_0^t e^{at'}\|w_\lambda(t')\|_{B^{s+1}_{p,1}}dt'+C\int_0^t e^{at'}\|g_\lambda(t')\|_{B^s_{p,1}}dt'.\end{aligned}$$

From this estimate, we get that

$$\begin{aligned}\|e^{at}\tau_\lambda(t)\|_{L^\infty_T(B^s_{p,1})}\leq{}&\|\tau_0\|_{B^s_{p,1}}+\frac{C}{\lambda}\|\tau_\lambda\|_{L^\infty_T(B^s_{p,1})}\\&+C\mu_2\int_0^T e^{at}\|w_\lambda(t)\|_{B^{s+1}_{p,1}}dt+C\int_0^T e^{at}\|g_\lambda(t)\|_{B^s_{p,1}}dt.\end{aligned}$$

So, if λ is large enough, we obtain

$$\|e^{at}\tau_\lambda(t)\|_{L^\infty_T(B^s_{p,1})}\leq 2\|\tau_0\|_{B^s_{p,1}}+C\mu_2\int_0^T e^{at}\|w_\lambda(t)\|_{B^{s+1}_{p,1}}dt+C\int_0^T e^{at}\|g_\lambda(t)\|_{B^s_{p,1}}dt.\tag{5.6}$$

In particular, this implies that

$$\|\tau_\lambda(t)\|_{L^1_T(B^s_{p,1})}\leq C\frac{1-e^{aT}}{a}\Big(\|\tau_0\|_{B^s_{p,1}}+\mu_2\|w_\lambda\|_{L^1_T(B^{s+1}_{p,1})}+\|g_\lambda\|_{L^1_T(B^s_{p,1})}\Big).\tag{5.7}$$

Classical estimates about the heat equation (see, for instance, [6]) give

$$\|\Delta_q w(t)\|_{L^p}\leq C\|\Delta_q w_0\|_{L^p}e^{-C^{-1}\nu 2^{2q}t}+C\int_0^t e^{-C^{-1}\nu 2^{2q}(t-t')}\|\Delta_q\widetilde{f}(t')\|_{L^p}dt'\tag{5.8}$$

with $\widetilde{f}\overset{\text{def}}{=}f+\mu_1\nabla\cdot\tau$. Then multiplying by $2^{qs}\exp\Big(-\lambda\int_0^t\|\nabla v(t')\|_{B^{s_c}_{p,1}}dt'\Big)$, taking the sum over q and integrating in time, we get

$$\|w_\lambda\|_{L^1_T(B^{s+1}_{p,1})}\leq\frac{C}{\nu}\left(\sum_q\|\Delta_q w_0\|_{L^p}(1-e^{-C^{-1}\nu 2^{2q}T})+\|\widetilde{f}_\lambda\|_{L^1_T(B^{s-1}_{p,1})}\right).$$

By the definition of $\widetilde{f}$, we obtain, applying the estimate (5.7),

$$\begin{aligned}\|w_\lambda\|_{L^1_T(B^{s+1}_{p,1})}\leq\frac{C}{\nu}\Bigg(&\sum_q\|\Delta_q w_0\|_{L^p}(1-e^{-C^{-1}\nu 2^{2q}T})+\|\widetilde{f}_\lambda\|_{L^1_T(B^{s-1}_{p,1})}\\&+\mu_1\frac{1-e^{aT}}{a}\Big(\|\tau_0\|_{B^s_{p,1}}+\mu_2\|w_\lambda\|_{L^1_T(B^{s+1}_{p,1})}+\|g_\lambda\|_{L^1_T(B^s_{p,1})}\Big)\Bigg).\end{aligned}$$

Then the condition (1.8) gives the estimate (5.2). To prove the inequality (5.3), let us go back to the estimate (5.8). Multiplying by $2^{q(s-1)}$ and taking the supremum in time gives

$$(5.9)\quad 2^{q(s-1)}\|\Delta_q w_\lambda\|_{L^\infty_T(L^p)} \leq C2^{q(s-1)}\|\Delta_q w_0\|_{L^p} + \int_0^T 2^{q(s-1)}\|\Delta_q \widetilde{f}_\lambda(t)\|_{L^p}dt.$$

Summing over q and using that

$$\|w_\lambda\|_{L^\infty_T(B^{s-1}_{p,1})} \leq \sum_q 2^{q(s-1)}\|\Delta_q w_\lambda\|_{L^\infty_T(L^p)},$$

we claim that

$$\|w_\lambda\|_{L^\infty_T(B^{s-1}_{p,1})} \leq C(\|w_0\|_{B^{s-1}_{p,1}} + \|\widetilde{f}_\lambda\|_{L^1_T(B^s_{p,1})}).$$

Then using the estimates (5.2) and (5.7) gives the inequality (5.3). To obtain the inequality (5.4), it is enough to plug the estimate (5.2) into (5.6).

Remark. In fact, we proved a better estimate which is

$$\sum_q 2^{q(s-1)}\|\Delta_q w_\lambda\|_{L^\infty_T(L^p)} \leq C(\|w_0\|_{B^{s-1}_{p,1}} + \|\widetilde{f}_\lambda\|_{L^1_T(B^s_{p,1})}).$$

However, we have to prove the inequality (5.5). The law of product in Besov spaces implies that

$$\|Q(\tau(t),\nabla v(t))\|_{B^s_{p,1}} \leq C\|\tau(t)\|_{B^s_{p,1}}\|\nabla v(t)\|_{B^{s_c}_{p,1}}$$

because s is in the interval $]-s_c, s_c+1]$. Thus, we have

$$\|\Delta_q Q(\tau(t),\nabla v(t))\|_{L^p} \leq Cc_q(t)2^{-qs}\|\tau(t)\|_{B^s_{p,1}}\|\nabla v(t)\|_{B^{s_c}_{p,1}}.$$

Then let us observe that, in [9], it is proved that

$$\|[v(t)\cdot\nabla, \Delta_q]\tau(t)\|_{L^p} \leq Cc_q(t)2^{-qs}\|\nabla v(t)\|_{B^{s_c}_{p,1}}\|\tau(t)\|_{B^s_{p,1}}.$$

The estimate (5.5) is proved and so is the lemma.

Now let us prove Theorem 1.6. As $\mathcal{D}$ (the space of smooth compactly supported functions) is dense in $B^{s_c-1}_{p,1}\times B^{s_c}_{p,1}$, let us consider a sequence $(v_{n,0},\tau_{n,0})$ of $\mathcal{D}$ which converges to (v_0,τ_0) in the Banach space $B^{s_c-1}_{p,1}\times B^{s_c}_{p,1}$ and such that for all n, we have $||v_{n,0}||_{B^{s_c}_{p,1}} \leq ||v_0||_{B^{s_c}_{p,1}}$ and $||\tau_{n,0}||_{B^{s_c-1}_{p,1}} \leq \|\tau_0\|_{B^{s_c-1}_{p,1}}$.

Theorem 1.1 claims that a smooth solution (v_n,τ_n) exists on a time interval $[0,T_n[$. Let us define

$$v_{n,L} = e^{\nu t\Delta}v_{n,0} \quad\text{and}\quad w_n = v_n - v_{n,L}.$$

Now let us apply the above Lemma 5.1 with $s = s_c$ and

$$f_n = \mathcal{Q}(w_n,w_n) + 2\mathcal{Q}(v_{n,L},w_n) + \mathcal{Q}(v_{n,L},v_{n,L}),\quad v = v_n,\quad g = 0.$$

Let us define

$$W_{n,\lambda}(T) = \nu||w_{n,\lambda}||_{L^1_T(B^{s_c+1}_{p,1})} + ||w_{n,\lambda}||_{L^\infty_T(B^{s_c-1}_{p,1})},$$

where we recall that

$$w_\lambda(t) = w(t)\exp\left(-\lambda\int_0^t \|\nabla v(t')\|_{B^{s_c}_{p,1}}dt'\right).$$

Then we get, for any $T \leq \min(T_{VE}, T_n)$,

$$W_{n,\lambda}(T) \leq C\left(||f_{n,\lambda}||_{L^1_T(B^{s_c-1}_{p,1})} + \frac{\mu_1}{a}(1-e^{-a\ T})||\tau_{n,0}||_{B^{s_c}_{p,1}}\right).$$

Let us estimate $||f_{n,\lambda}||_{L^1_T(B^{s_c-1}_{p,1})}$. As $B^{s_c}_{p,1}$ is an algebra,

$$||Q(a,b)||_{L^1_T(B^{s_c-1}_{p,1})} \leq C||a||_{L^2_T(B^{s_c}_{p,1})}\ ||b||_{L^2_T(B^{s_c}_{p,1})}.$$

Then using classical interpolation results, we get

$$||a||_{L^2_T(B^{s_c}_{p,1})} \leq ||a||^{1/2}_{L^1_T(B^{s_c+1}_{p,1})}||a||^{1/2}_{L^\infty_T(B^{s_c-1}_{p,1})}.$$

We thus infer that

$$||f_{n,\lambda}||_{L^1_T(B^{s_c-1}_{p,1})} \leq C\left(W^2_{n,\lambda} + W_{n,\lambda}||v_{n,L}||_{L^2_T(B^{s_c}_{p,1})} + ||v_{n,L}||^2_{L^2_T(B^{s_c}_{p,1})}\right).$$

Using Lemma 5.1, we get that

$$W_{n,\lambda}(T) \leq C\left(||v_{n,L}||^2_{L^2_T(B^{s_c}_{p,1})} + W^2_{n,\lambda}(T) + \frac{\mu_1}{a}(1-e^{-a\ T})||\tau_{n,0}||_{B^{s_c}_{p,1}}\right).$$

Let $T \leq \min\{T_{(v_{n,0},\tau_{n,0})}, T_{VE}, T_n\}$, where $T = T_{(v_{n,0},\tau_{n,0})}$ is such that

$$\frac{C\mu_1}{a\nu}(1-e^{-a\ T})||\tau_{n,0}||_{B^{s_c}_{p,1}} + \frac{C}{\nu}\left(\sum_q 2^{q(s-1)}\|\Delta_q v_{n,0}\|_{L^p}(1-e^{C^{-1}\nu 2^{2q}T})\right) \leq \frac{1}{8C}.$$

We can see easily that $T_{(v_{n,0},\tau_{n,0})}$ goes to $T_{(v_0,\tau_0)}$, when n goes to ∞. Next we recall that (see, for instance, [6])

$$||v||_{L^p_T(B^{s_c-1+\frac{2}{p}}_{p,1})} \leq \sum_q 2^{q(s-1)}\|\Delta_q v\|_{L^p}\left(\frac{1-e^{C^{-1}\nu 2^{2q}T}}{\nu}\right)^{1/p}.$$

So we get

$$W_{n,\lambda}(T) \leq CW^2_{n,\lambda}(T) + \frac{1}{5C}$$

and, since $W_{n,\lambda}(0) = 0$ and that $W_{n,\lambda}(t)$ is continuous in t, we deduce that for all T

$$W_{n,\lambda}(T) \leq \frac{1}{2C}.$$

This can be rewritten as follows:

$$\nu||w_n||_{L^1_T(B^{s_c+1}_{p,1})} + ||w_n||_{L^\infty_T(B^{s_c-1}_{p,1})} \leq \frac{1}{2C}\exp\left(\lambda||v_{n,L}||_{L^1_T(B^{s_c+1}_{p,1})} + \lambda||w_n||_{L^1_T(B^{s_c+1}_{p,1})}\right).$$

If C is chosen big enough, we get

$$||w_n||_{L^1_T(B^{s_c+1}_{p,1})} \leq \frac{1}{3}\exp(||w_n||_{L^1_T(B^{s_c+1}_{p,1})}).$$

Then using the following lemma, we conclude easily that $||w_n||_{L^1_T(B^{s_c+1}_{p,1})} \leq 1$.

Lemma 5.2. *If $f(t)$ is a continuous function satisfying for any t in $[0,T]$*

$$f(t) \leq \eta e^{f(t)}$$

with $\eta < \frac{1}{e}$ and $f(0)=0$, then we have for all $t \in [0,T]$,

$$f(t) \leq e\eta.$$

Then we deduce that w_n is bounded in $L^\infty_T(B^{s_c-1}_{p,1}) \cap L^1_T(B^{s_c+1}_{p,1})$. Using Lemma 5.1 for τ, we get that τ_n is bounded in $L^\infty_T(B^{s_c}_{p,1})$. The explosion condition in Sobolev spaces and the fact that $||\nabla v||_{L^\infty} \leq ||v||_{B^{s_c+1}_{p,1}}$ and $||\tau||_{L^\infty} \leq ||\tau||_{B^{s_c}_{p,1}}$ show that $T_n \geq \min\{T_{(v_{n,0},\tau_{n,0})}, T_{VE}\}$ and hence, one can take $T = \min\{T_{(v_0,\tau_0)}, T_{VE}\}$.

Now let us prove the uniqueness part. We recall that we assume here that p is in the interval $[1,2d]$. This means that $s_c \geq \frac{1}{2}$. Let us consider (v,τ) a solution of (VE) in $L^\infty_T(B^{s_c-1}_{p,1}) \cap L^1_T(B^{s_c+1}_{p,1}) \times L^\infty_T(B^{s_c}_{p,1})$. It is obvious that

$$\partial_t v \in L^1_T(B^{s_c-1}_{p,1}) \quad \text{and that} \quad \partial_t \tau \in L^1_T(B^{s_c-1}_{p,1}).$$

So we get $(v-v_0, \tau-\tau_0) \in C([0,T]; B^{s_c-1}_{p,1} \times B^{s_c-1}_{p,1})$. Let us consider two solutions (v_j,τ_j) of (VE) associated to the same initial data. The difference $(w,\theta) \overset{\text{def}}{=} (v_1-v_2, \tau_1-\tau_2)$ is in $C([0,T]; B^{s_c-1}_{p,1} \times B^{s_c-1}_{p,1})$ and satisfies

$$\left\{\begin{array}{rcl} \partial_t w - \nu\Delta w & = & P\mu_1\nabla\tau + 2\mathcal{Q}(v_1+v_2,w), \\ \partial_t\theta + v_1\cdot\nabla\theta + a\theta & = & \mu_2 Dw - Q(\theta,\nabla v_1) - Q(\tau_2,\nabla w) - w\cdot\nabla\tau_2, \\ (w,\theta)_{|t=0} & = & (0,0). \end{array}\right.$$

Applying Lemma 5.1 with $s = s_c - 1$, we get

$$\nu||w_\lambda||_{L^1_T(B^{s_c}_{p,1})} + ||w_\lambda||_{L^\infty_T(B^{s_c-2}_{p,1})} + ||\theta_\lambda||_{L^\infty_T(B^{s_c-1}_{p,1})} \leq C\left(||f_\lambda||_{L^1_T(B^{s_c-2}_{p,1})} + ||g_\lambda||_{L^1_T(B^{s_c-1}_{p,1})}\right),$$

where

$$f_\lambda = 2\mathcal{Q}(v_1+v_2,w) \quad \text{and} \quad g_\lambda = 2\mathcal{Q}(v_1+v_2,w) = -Q(\theta,\nabla v_1) - Q(\tau_2,\nabla w) - w\cdot\nabla\tau_2.$$

Since we are looking for uniqueness, we can forget the λ. In what follows the constant C will denote $C\exp(-\lambda\int_0^T \|\nabla v_1(t')\|_{B^{s_c}_{p,1}} dt')$. Using classical results for products in Besov spaces and the fact that $s_c + (s_c-1) \geq 0$, we get

$$\begin{array}{rcl} ||f||_{L^1_T(B^{s_c-2}_{p,1})} & \leq & C\left(||v_1||_{L^2_T(B^{s_c}_{p,1})} + ||v_2||_{L^2_T(B^{s_c}_{p,1})}\right)||w||_{L^2_T(B^{s_c-1}_{p,1})} \\ & \leq & C\left(||v_1||_{L^2_T(B^{s_c}_{p,1})} + ||v_2||_{L^2_T(B^{s_c}_{p,1})}\right)||w||^{\frac{1}{2}}_{L^1_T(B^{s_c}_{p,1})}||w||^{\frac{1}{2}}_{L^\infty_T(B^{s_c-2}_{p,1})} \end{array}$$

and

$$||g||_{L^1_T(B^{s_c-1}_{p,1})} \leq C\left(\|v_1\|_{L^1_T(B^{s_c+1}_{p,1})}||\theta||_{L^\infty_T(B^{s_c-1}_{p,1})} + ||\tau_2||_{L^\infty_T(B^{s_c}_{p,1})}||w||_{L^1_T(B^{s_c}_{p,1})}\right).$$

So, uniqueness is proved by application of the Gronwall lemma.

Finally, we prove Theorem 1.9. Rewriting (2.5) in the framework of Besov spaces, we get for $q \geq 0$

$$\begin{aligned}V_q(t) &\overset{\text{def}}{=} \frac{d}{dt}(\mu_2\|\Delta_q v(t)\|^2_{L^2} + \mu_1\|\Delta_q\tau\|^2_{L^2}) + 2\mu_2\nu\|\nabla\Delta_q v_n(t)\|^2_{L^2} + 2\mu_1 a\|\Delta_q\tau_n\|^2_{L^2}\\ &\leq Cc_q(t)2^{-qs}\biggl(\|\nabla v(t)\|_{B^s_{p,1}}\mu_2\|v(t)\|_{B^s_{p,1}}\|\Delta_q v(t)\|_{L^2}\\ &\qquad\qquad + \|\nabla v(t)\|_{B^s_{p,1}}\mu_1\|\tau(t)\|_{B^s_{p,1}}\|\Delta_q\tau(t)\|_{L^2}\\ &\qquad\qquad + \|\tau(t)\|_{B^s_{p,1}}\mu_1\|\nabla v(t)\|_{B^s_{p,1}}\|\Delta_q\tau(t)\|_{L^2}\biggr),\end{aligned}$$

where the series $c_q(t)$ is now such that $\sum_q c_q(t) = 1$. Then using that

$$(\mu_2\|\Delta_q v(t)\|^2_{L^2} + \mu_1\|\Delta_q\tau\|^2_{L^2}) \geq \frac{1}{2}(\sqrt{\mu}_2\|\Delta_q v(t)\|_{L^2} + \sqrt{\mu}_1\|\Delta_q\tau\|_{L^2})^2$$

and that

$$\begin{aligned}2\mu_2 2^{2q}\nu\|\Delta_q v(t)\|^2_{L^2} + 2\mu_1 a\|\Delta_q\tau\|^2_{L^2} &\geq (2^q\sqrt{\nu\mu_2}\|\Delta_q v(t)\|_{L^2} + \sqrt{a\mu_1}\|\Delta_q\tau\|_{L^2})\\ &\quad\times\min(\sqrt{\nu},\sqrt{a})(\sqrt{\mu}_2\|\Delta_q v(t)\|_{L^2} + \sqrt{\mu}_1\|\Delta_q\tau\|_{L^2}),\end{aligned}$$

we get

$$\begin{aligned}&\frac{d}{dt}(\sqrt{\mu}_2\|\Delta_q v(t)\|_{L^2} + \sqrt{\mu}_1\|\Delta_q\tau\|_{L^2}) + \min(\sqrt{\nu},\sqrt{a})(2^q\sqrt{\nu\mu_2}\|\Delta_q v(t)\|_{L^2} + \sqrt{a\mu_1}\|\Delta_q\tau\|_{L^2})\\ &\leq Cc_q(t)2^{-qs}\biggl(\|\nabla v(t)\|_{B^s_{p,1}}\sqrt{\mu_2}\|v(t)\|_{B^s_{p,1}}\\ &\qquad + \|\nabla v(t)\|_{B^s}\sqrt{\mu}_1\|\tau(t)\|_{B^s_{p,1}} + \|\tau(t)\|_{B^s_{p,1}}\sqrt{\mu}_1\|\nabla v(t)\|_{B^s_{p,1}}\biggr).\end{aligned}$$

Then multiplying by 2^{qs} and summing in q, we get

$$\begin{aligned}\frac{d}{dt}\biggl(\sqrt{\mu}_2\|v\|_{B^s_{p,1}} + \sqrt{\mu}_1\|\tau\|_{B^s_{p,1}}\biggr) &+ \min(\sqrt{\nu},\sqrt{a})(\sqrt{\nu\mu_2}\|v\|_{B^{s+1}_{p,1}} + \sqrt{a\mu_1}\|\tau\|_{B^s_{p,1}})\\ &\leq C\|v(t)\|_{B^{s+1}_{p,1}}\biggl(\sqrt{\mu_2}\|v(t)\|_{B^s_{p,1}} + \sqrt{\mu}_1\|\tau\|_{B^s_{p,1}}\biggr).\end{aligned}$$

Thus, if

$$(\sqrt{\mu}_2\|v_0\|_{B^s_{p,1}} + \sqrt{\mu}_1\|\tau_0\|_{B^s_{p,1}}) \leq c\min(\sqrt{\nu},\sqrt{a})\sqrt{\nu\mu_2},$$

we get the global existence.

REFERENCES

[1] H. Bahouri and J.-Y. Chemin, *Equations de transport relatives à des champs de vecteurs non-lipschitziens et méchanique des fluides*, Arch. Rational Mech. Anal., 127 (1994), pp. 159–182.

[2] J. Beale, T. Kato, and A. Majda, *Remarks on the breakdown of smoothness for the* 3-*D Euler equations*, Comm. Math. Phys., 94 (1984), pp. 61–66.

[3] J.-M. Bony, *Calcul symbolique et propagation des singularités pour les équations aux dérivées partielles non linéaires*, Ann. Sci. École Norm. Sup. (4), 14 (1981), pp. 209–246.

[4] M. Cannone, *Ondelettes, paraproduits et Navier–Stokes*, Diderot Éditeur, Paris, 1995.

[5] J.-Y. Chemin, *Perfect Incompressible Fluids*, Oxford University Press, New York, 1998.

[6] J.-Y. Chemin, *Théorèmes d'unicité pour le système de Navier–Stokes tridimensionnel*, J. Anal. Math., 77 (1999), pp. 27–50.

[7] J.-Y. Chemin and N. Lerner, *Flot de champs de vecteurs non-lipschitziens et équations de Navier–Stokes*, J. Differential Equations, 121 (1995), pp. 314–328.

[8] F. Colombini and N. Lerner, *Hyperbolic operators with non-Lipschitz coefficients*, Duke Math. J., 77 (1995), pp. 657–698.

[9] R. Danchin, *Poches de tourbillon visqueuses*, J. Math. Pures Appl. (9), 76 (1997), pp. 609–647.

[10] R. Danchin and B. Desjardins, *Existence of solutions for compressible fluid models of Korteweg type*, Ann. Inst. H. Poincaré Anal. Non Linéaire, to appear.

[11] B. Desjardins, *Linear transport equations with values in Sobolev spaces and application to the Navier–Stokes equations*, Differential Integral Equations, 10 (1997), pp. 577–586.

[12] H. Fujita and T. Kato, *On the Navier-Stokes initial value problem* I, Arch. Rational Mech. Anal., 16 (1964), pp. 269–315.

[13] C. Guillopé and J.-C. Saut, *Mathematical problems arising in differential models for viscoelastic fluids*, in Mathematical Topics in Fluid Mechanics, Lisbon, 1991, Pitman Res. Notes in Math. 274, Longman, Harlow, UK, 1992, pp. 64–92.

[14] C. Guillopé and J.-C. Saut, *Existence results for the flow of viscoelastic fluids with a differential constitutive law*, Nonlinear Anal., 15 (1990), pp. 849–869.

[15] C. Guillopé and J.-C. Saut, *Global existence and one-dimensional nonlinear stability of shearing motions of viscoelastic fluids of Oldroyd type*, RAIRO Modél. Math. Anal. Numér., 24 (1990), pp. 369–401.

[16] E. Fernandez-Cara, F. Guillén, and R.R. Ortega, *Existence et unicité de solution forte locale en temps pour des fluides non newtoniens de type Oldroyd (version L^s–L^r)*, C. R. Acad. Sci. Paris Sér. I Math., 319 (1994), pp. 411–416.

[17] P.-L. Lions and N. Masmoudi, *Global solutions for some Oldroyd models of non-Newtonian flows*, Chinese Ann. Math. Ser. B, 21 (2000), pp. 131–146.

[18] Y. Meyer, *Ondelettes et opérateurs tome* 2, Herman, 1991.

[19] J.G. Oldroyd, *Non-Newtonian effects in steady motion of some idealized elastico-viscous liquids*, Proc. Roy. Soc. London Ser. A, 245 (1958), pp. 278–297.

[20] M. Renardy, *Existence of slow steady flows of viscoelastic fluids with differential constitutive equations*, Z. Angew. Math. Mech., 65 (1985), pp. 449–451.

[21] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland, Amsterdam, 1978.

SIAM J. MATH. ANAL.
Vol. 33, No. 1, pp. 113–152

ESTIMATES FOR PERIODIC AND DIRICHLET EIGENVALUES OF THE SCHRÖDINGER OPERATOR[*]

T. KAPPELER[†] AND B. MITYAGIN[‡]

Abstract. Consider the Schrödinger equation $-y'' + Vy = \lambda y$ for a complex-valued potential V of period 1 in the weighted Sobolev space H^w of 2-periodic functions $f : \mathbb{R} \to \mathbb{C}$,

$$H^w \equiv H^w_{\mathbb{C}} := \left\{ f(x) = \sum_{k=-\infty}^{\infty} \hat{f}(k) e^{i\pi kx} | \; \|f\|_w < \infty \right\},$$

where

$$\|f\|_w := \left(2 \sum_k w(k)^2 \; |\hat{f}(k)|^2 \right)^{1/2}$$

and $w = (w(k))_{k \in \mathbb{Z}}$ denotes a symmetric, submultiplicative weight sequence. Denote by $\lambda_n = \lambda_n(V)$ $(n \geq 0)$ the periodic eigenvalues of $-\frac{d^2}{dx^2} + V$ when considered on the interval $[0, 2]$, listed in such a way that $\lambda_{2n}, \lambda_{2n-1} = n^2\pi^2 + 0(1)$, and denote by $\mu_n = \mu_n(V)$ $(n \geq 1)$ the Dirichlet eigenvalues of $-\frac{d^2}{dx^2} + V$ considered on $[0, 1]$, listed in such a way that $\mu_n = n^2\pi^2 + 0(1)$.

THEOREM. *There exist (absolute) constants $K_1, K_2 > 0$, so that for any 1-periodic potential V in H^w,*

$$\sum_{n \geq N} w(2n)^2 |\lambda_{2n} - \lambda_{2n-1}|^2 \leq K_1(1 + \|V\|_w)^{K_2}$$

and

$$\sum_{n \geq N} w(2n)^2 |\mu_n - \lambda_{2n}|^2 \leq K_1(1 + \|V\|_w)^{K_2},$$

where $N := K_1(1 + \|V\|_w)^2$.

Key words. Schrödinger operators, periodic and Dirichlet eigenvalues, estimates on gap lengths

AMS subject classifications. 58F19, 58F07, 35Q35

PII. S0036141099365753

1. Introduction.

1.1. Summary of the results.

Consider the Schrödinger equation on the interval $[0, 2]$,

$$-y'' + Vy = \lambda y, \tag{1.1}$$

where V is a complex-valued periodic potential of *period* 1 in the weighted Sobolev space of 2-periodic functions,

$$H^w \equiv H^w_{\mathbb{C}} := \left\{ f(x) = \sum_k \hat{f}(k) e^{i\pi kx} \; | \; \|f\|_w < \infty \right\}$$

[*] Received by the editors November 29, 1999; accepted for publication (in revised form) November 28, 2000; published electronically April 24, 2001.
http://www.siam.org/journals/sima/33-1/36575.html

[†] Institut für Mathematik, Universität Zürich, 8057 Zürich, Switzerland (tk@math.unizh.ch).

[‡] Department of Mathematics, Ohio State University, Columbus, OH 43210 (borismit@math.ohio-state.edu).

with

$$\|f\|_w := \left(2 \sum_k w(k)^2 \; |\hat{f}(k)|^2 \right)^{1/2}$$

and $w = (w(k))_{k\in\mathbb{Z}}$ with $w(k) \geq 1 \; \forall k \in \mathbb{Z}$ is a symmetric weight $(w(k) = w(-k) \; \forall k \in \mathbb{Z})$ which is submultiplicative,

$$w(k+j) \leq w(k)w(j) \; \forall k, j \in \mathbb{Z}.$$

As an example of a submultiplicative weight we mention the Abel–Sobolev weight $w_{a,b}(k) := (1+|k|)^a e^{b|k|}$ with $a \geq 0$ and $b \geq 0$. An element $f \in H^{w_{a,b}}$ can be viewed as a complex-valued function $F(z) = \sum_{k\in\mathbb{Z}} \hat{f}(k)e^{i\pi kz}$, $z = x+iy$, analytic in the strip $|y| < \frac{b}{\pi}$ and such that $F(x + i\frac{b}{\pi})$ as well as $F(x - i\frac{b}{\pi})$ are in the Sobolev space H^a, defined by the weight $w(k) := (1+|k|)^a$. More generally, $w(k) := (1+|k|)^a e^{b|k|^\alpha}$ is a submultiplicative weight for $0 \leq \alpha \leq 1, a \geq 0$ and $b \geq 0$; a function $f \in H^w$ is a function of Gevrey class.

The spectrum $spec_{Per}(L)$ of the operator $L := -\frac{d^2}{dx^2} + V$, when considered on the interval $[0, 2]$ and with periodic boundary conditions, is discrete and is a sequence $\lambda_n = \lambda_n(V)$ $(n \geq 0)$ with the property that $\text{Re}\lambda_n \to +\infty$ for $n \to \infty$. Here, the eigenvalues λ_n are enumerated with their algebraic multiplicities and ordered so that

$$\text{Re}\lambda_n < \text{Re}\lambda_{n+1} \qquad \text{or} \qquad \text{Re}\lambda_n = \text{Re}\lambda_{n+1} \text{ and } \text{Im}\lambda_n \leq \text{Im}\lambda_{n+1}.$$

Notice that adding a constant to the potential V results in a shift of the eigenvalues by the same constant. Hence we restrict ourselves—without loss of generality—to potentials V of mean zero and introduce the subspace $H_0^w \subseteq H^w$,

$$H_0^w := \left\{ f \in H_{\mathbb{C}}^w \;\middle|\; \int_0^2 f(x)dx = 0 \right\}.$$

For the weight $w \equiv 1$, the spaces H_0^w and H^w are also denoted by L_0^2 and L^2, respectively. For n sufficiently large (cf. Lemma 1.4 in section 1.2 for a reminder), the eigenvalues come in pairs $\{\lambda_{2n}, \lambda_{2n-1}\}$, i.e., λ_{2n} and λ_{2n-1} are close to each other and separated from the rest of $spec_{per}(L)$ by a distance of size n.

In section 2 of this paper, we prove the following theorem.

Theorem 1.1. *There exist (absolute) constants $K_1, K_2 > 0$ so that for any 1-periodic potential V in H_0^w*

$$\sum_{n\geq N} w(2n)^2 \; |\lambda_{2n} - \lambda_{2n-1}|^2 \leq K_1(1 + \|V\|_w)^{K_2}, \tag{1.2}$$

where $N := K_1(1 + \|V\|_w)^2$.

(See Proposition 2.16 and section 2.8 for details.)

In the next theorem we state the main two terms in the asymptotics of the sequence of gap lengths, $\gamma_n := \lambda_{2n} - \lambda_{2n-1}$ as $n \to \infty$. For this purpose introduce

$$\rho(n) := \hat{V}(2n) + \frac{1}{\pi^2} \sum_j \frac{\hat{V}(n-j)}{n-j} \frac{\hat{V}(n+j)}{n+j}. \tag{1.3}$$

Notice that the last term in (1.3) is a convolution and well defined as $\hat{V}(0) = 0$.

Theorem 1.2. *There exist (absolute) constants $K_3, K_4 > 0$ so that for any 1-periodic potential V in H_0^w*

$$\sum_{n\geq N}(1+|n|)^2 w(2n)^2 \min_{\pm} \mid (\lambda_{2n}-\lambda_{2n-1}) \pm 2\sqrt{\rho(n)\rho(-n)} \mid^2 \leq K_3(1+\|V\|_w)^{K_4}, \tag{1.4}$$

where $N := K_3(1+\|V\|_w)^2$.

(See Theorem 2.20 and section 2.9 for further details.)

In our previous paper [8], we obtained estimate (1.2) and a weaker form of estimate (1.4) for the Abel–Sobolev weights

$$w_{a,b} := (1+|k|)^a e^{b|k|}, \quad (a \geq 0, b \geq 0),$$

using a Fourier approach. By a refined analysis we obtain in section 2 of the present paper estimates (1.2) for general submultiplicative weights and a two-terms asymptotic (1.3)–(1.4) for the gap lengths. It turns out that submultiplicative weights provide the right setup for applications to a KAM theorem for the Korteweg–deVries equation (cf. [2]), as will be shown in a subsequent paper. Further, we present in the present paper an analysis of the Riesz spaces together with estimates for the Dirichlet eigenvalues. Let us explain this in more detail.

In section 3, we analyze the Riesz spaces E_n (n sufficiently large), i.e., the images of the Riesz projectors defined by a circle of appropriate size around $n^2\pi^2$ as contour and the operator $L := \frac{d^2}{dx^2}+V$, considered on $[0,1]$ with periodic (for n even) or antiperiodic (for n odd) boundary conditions (cf. (1.16)). We study the structure of L by computing the matrix representation of the restriction of $L-\lambda_{2n}$ to E_n with respect to an orthonormal basis f_n, φ_n, where f_n is a periodic or antiperiodic eigenfunction in E_n. Moreover, we estimate the entries of this matrix which will be important for estimates of the Dirichlet eigenvalues (cf. Theorem 3.5 and Proposition 3.6).

In section 4 we obtain estimates for the Dirichlet eigenvalues $\mu_n(V)$ ($n \geq 1$) of the operator $-\frac{d^2}{dx^2}+V$, considered on the interval $[0,1]$. The eigenvalues $\mu_n \equiv \mu_n(V)$ are ordered in such a way that

$$\mathrm{Re}\mu_n < \mathrm{Re}\mu_{n+1} \qquad \text{or} \qquad \mathrm{Re}\mu_n = \mathrm{Re}\mu_{n+1} \text{ and } \mathrm{Im}\mu_n \leq \mathrm{Im}\mu_{n+1}. \tag{1.5}$$

Theorem 1.3. *There exist (absolute) constants $K_5, K_6 > 0$ so that for any 1-periodic potential V in H_0^w*

$$\sum_{n\geq N} w(2n)^2|\mu_n - \lambda_{2n}|^2 \leq K_5(1+\|V\|_w)^{K_6}, \tag{1.6}$$

where $N := K_5(1+\|V\|_w)^{K_6}$.

It turns out that by the methods used to prove Theorem 1.3, one can obtain similar results for the eigenvalues of L_{bc}, where L_{bc} is the operator L with boundary conditions bc from a special class $\mathcal{B}$. In section 5, this class is defined and the spectrum of the operators L_{bc} is analyzed.

It is well known that the decay of the gap lengths $\gamma_n := \lambda_{2n} - \lambda_{2n-1}$, associated to $spec_{Per}(L)$, depends on the smoothness properties of V (cf., e.g., [7], [13], [20]). In particular Marčenko [13] obtains polynomial decay of the gap lengths in terms of the Sobolev class of the potential and Trubowitz [20] proves exponential decay for real analytic potentials. Conversely, the question of smoothness of an L_2-potential in terms of the decay of the gap lengths has been addressed as well, mainly for real-valued potentials (cf. [13], [15], [20]) but more recently also for complex-valued potentials. It turns out that for complex-valued potentials, the decay of the gap lengths does not suffice to determine the smoothness: Sansuc and Tkachenko [19] proved that a periodic complex-valued potential $V \in L_0^2$ belongs to the Sobolev space H_0^N iff the following two conditions are satisfied:

$$\sum_{n\geq 1}(1+|n|)^{2N}\,|\lambda_{2n}-\lambda_{2n-1}|^2 < \infty \ ; \quad \sum_{n\geq 1}(1+|n|)^{2N}\,|\mu_n-\lambda_{2n}|^2 < \infty,$$

where, as above, $(\mu_n)_{n\geq 1}$ denote Dirichlet eigenvalues.

The condition of the weight sequence $(w(n))_{n\in\mathbb{Z}}$ to be submultiplicative could be seen as purely technical and convenient in the proofs of the inequalities stated in the theorems above, but it may be too restrictive for results like Theorem 1.1. Moreover, the submultiplicativity implies that

$$\lim_{n\to\infty}\frac{\log w(n)}{n} = \omega_* < \infty.$$

Thus, for $\omega > \omega_*$,

$$w(n) \leq C_\omega e^{\omega|n|} \quad \forall n \in \mathbb{Z} \tag{1.7}$$

for some constant $C_\omega > 0$ and $w(n)$ cannot grow faster than an exponential function. Notice, however, that the slightest violation of the growth restriction (1.7) gives a weight sequence which does not have the property stated by Theorem 1.1. This follows from Harrell's and Grigis's analysis of the gap lengths for (real) polynomial potentials. If V is a Mathieu potential

$$V(x) = t\cos(2\cdot 2\pi x) \quad 0 \leq x \leq 1, \tag{1.8}$$

then Harrell [6] (cf. [1]) proved that the gap lengths γ_n satisfy the asymptotic estimates (cf. [4, formula (1.8)])

$$\gamma_n = \frac{t^n}{8^{n-1}((n-1)!)^2}\left(1+0\left(\frac{1}{n^2}\right)\right),$$

and therefore, for some a, depending on t,

$$\gamma_n > e^{-2n\log n + an}. \tag{1.9}$$

Hence, if $w(n) := e^{b|n|\log|n|}$ with $b > 2$, the analogue of Theorem 1.1 does not hold. Indeed, we have

$$\|V\|_w^2 = \frac{\pi|t|^2}{2}w(4)^2 < \infty,$$

but (compare with (1.2))

$$\sum_{n\geq N} w(2n)^2|\gamma_n|^2 \geq \sum_{n\geq N} e^{4bn\log n}e^{-4n\log n+2an} = \infty.$$

A more refined analysis due to Grigis (see [4, Theorem 0.2]) shows that the above weight is bad with any $b > 0$, i.e., does not have the property stated by Theorem 1.1.

1.2. Preliminaries. General references on Schrödinger operators on the interval and Hill's operator can be found, e.g., in [11], [12], [18].

In this section we put together some well-known spectral properties of the operator $L := -\frac{d^2}{dx^2} + V$ in a form convenient for our further analysis. The following three lemmas are particular results in the general theory of nonselfadjoint boundary value problems developed by Keldysh [9], [10]. Many details can be found in [14], section 6 of chapter 1, in particular in subsection 6.3 (Lemmas 6.6 and 6.7) and 6.4 (p. 34); cf. also the appendix (pp. 215–219) where the paper [9] is translated into English.

Let us consider Dirichlet boundary conditions, $bc = Dir$, as well as periodic Per^+ and antiperiodic Per^- boundary conditions, $bc = Per^{\pm}$, i.e., for functions y in $H^2_{\mathbb{C}}[0,1]$,

(Dir) $\quad y(0) = 0\ ; \quad y(1) = 0;$

(Per^+) $\quad y(1) = y(0)\ ; \quad y'(1) = y'(0);$

(Per^-) $\quad y(1) = -y(0)\ ; \quad y'(1) = -y'(1).$

For $V \in L^2_{\mathbb{C}}[0,1]$ with $\int_0^1 V(x)dx = 0$ introduce the operator $L := D^2 + V$, where $D = \frac{1}{i}\frac{d}{dx}$. Given one of the above boundary conditions bc, denote by L_{bc} the closed operator in $L^2_{\mathbb{C}}[0,1]$ with domain $dom(L_{bc}) := \{f \in H^2_{\mathbb{C}}([0,1]) | f \text{ satisfies } bc\}$. Let $spec_{bc}(L) \equiv spec(L_{bc})$ be the spectrum of L_{bc}. For the potential $V \equiv 0$, i.e., $L = D^2$, $spec_{bc}(D^2)$ can be given explicitly,

$$spec_{Dir}(D^2) = \{k^2\pi^2 | k \geq 1\}, \tag{1.10}$$

$$spec_{Per^+}(D^2) = \{0\} \cup \{(2k)^2\pi^2, (2k)^2\pi^2 | k \geq 1\}, \tag{1.11}$$

$$spec_{Per^-}(D^2) = \{(2k-1)^2\pi^2, (2k-1)^2\pi^2 | k \geq 1\}. \tag{1.12}$$

For $r > 0$ and $k \in \mathbb{Z}_{\geq 0}$, let $\mathcal{D}(k) \equiv \mathcal{D}_r(k)$ be the open disc in $\mathbb{C}$ with center $k^2\pi^2$ and radius r

$$\mathcal{D}(k) := \{z \in \mathbb{C} \mid \quad |z - k^2\pi^2| < r\}$$

and, for $r_1, r_2 > 0, \mathcal{R} \equiv \mathcal{R}_{r_1,r_2}$ the open rectangle in $\mathbb{C}$

$$\mathcal{R} := \{x + iy \mid \quad -r_1 < x < r_2;\ |y| < r_2\}.$$

Denote by $\|V\|$ the L^2-norm of $V \in L^2_{\mathbb{C}}[0,2], \|V\| = (2\sum_k |\hat{V}(k)|^2)^{1/2}$.

Lemma 1.4. *There exist absolute constants $K_7 \geq 1$ and $K_8 \geq 1$ so that, for any given $M \geq 1$, boundary condition $bc \in \{Dir, Per^{\pm}\}$, $N \geq 2K_8(M+1)$, and 1-periodic potential $V \in L^2_{\mathbb{C}}[0,2]$ with $\|V\| \leq M$, the following holds:*

$$spec(L_{bc}) \subset \mathcal{R} \cup \bigcup_{k=N+1}^{\infty} \mathcal{D}(k), \tag{1.13}$$

where $\mathcal{D}(k) \equiv \mathcal{D}_r(k)$ with $r := K_8(M+1)$ and $\mathcal{R} = \mathcal{R}_{r_1,r_2}$ with $r_1 = K_7(1+M)^{4/3}$ and $r_2 = (N^2+N)\pi^2$.

We point out that $spec_{Per^+}(D^2) \cup spec_{Per^-}(D^2)$ (cf. (1.11) and (1.12)) is the spectrum $spec_{Per}(D^2)$ of the operator D^2 on $[0,2]$ with periodic boundary conditions. Obviously, for any 1-periodic potential V,

$$spec_{Per^+}(L) \cup spec_{Per^-}(L) \subseteq spec_{Per}(L).$$

For a real-valued potential the converse inclusion

$$spec_{Per}(L) \subseteq spec_{Per^+}(L) \cup spec_{Per^-}(L) \tag{1.14}$$

also holds, as one can see from an elementary application of Floquet theory. More generally, by a simple counting argument, Lemma 1.4 implies that (1.14) holds for complex-valued potentials.

The periodic eigenvalues of L on $[0,2]$ have been denoted by $(\lambda_n)_{n\geq 0}$ (cf. (1.7)). According to Lemma 1.4, the eigenvalues λ_{2n-1} and λ_{2n} are close to $n^2\pi^2$ for n sufficiently large. At certain occasions (cf., e.g., section 2.8), one of the two eigenvalues, either λ_{2n} or λ_{2n-1}, will satisfy a certain property, but it will not be possible to decide which of the two. For such a situation, it is convenient to introduce λ_n^+, λ_n^- as a different notation for the eigenvalues $\lambda_{2n}, \lambda_{2n-1}$,

$$\{\lambda_n^+, \lambda_n^-\} = \{\lambda_{2n}, \lambda_{2n-1}\}.$$

By Lemma 1.4, it follows that the Riesz projectors $P_* \equiv P_{*;bc}$ and $P_k \equiv P_{k;bc}$ are well defined for $\|V\| \leq M$,

$$P_* := \frac{1}{2\pi i}\int_{\partial\mathcal{R}} (z - L_{bc})^{-1} dz, \tag{1.15}$$

$$P_k := \frac{1}{2\pi i}\int_{\partial\mathcal{D}(k)} (z - L_{bc})^{-1} dz, \quad (k \geq N+1), \tag{1.16}$$

where the contours $\partial\mathcal{R}$ and $\partial\mathcal{D}(k)$ are counterclockwise oriented. Denote by $\|T\|_{\mathcal{L}(L^2)}$ the operator norm of a bounded linear operator $T : L^2_{\mathbb{C}}[0,1] \to L^2_{\mathbb{C}}[0,1]$.

Lemma 1.5. *There exist absolute constants K_9 and K_{10} so that under the same assumptions as in Lemma* 1.4*,*

$$\|P_*\|_{\mathcal{L}(L^2)} \leq K_9 \log(2+M), \tag{1.17}$$

$$\|P_k\|_{\mathcal{L}(L^2)} \leq K_{10}, \quad (k \geq N+1). \tag{1.18}$$

Further, for any $f \in L^2_{\mathbb{C}}[0,1]$,

$$f = P_* f + \sum_{k=N+1}^{\infty} P_k f, \tag{1.19}$$

where the series (1.19) *converges in L^2.*

Lemma 1.6. *There exists an absolute constant $K_{11} \geq 1$ so that under the same assumptions as in Lemma* 1.4*,*

$$\|(\lambda - L_{bc})^{-1}(Id - P_k)\|_{\mathcal{L}(L^2)} \leq K_{11}\frac{1}{k} \quad \forall \lambda \in \mathcal{D}_r(k), \quad \forall k \geq N+1. \tag{1.20}$$

2. Periodic eigenvalues.

2.1. Fourier block decomposition. Denote by L the Schrödinger operator $L := D^2 + V$, $D = \frac{1}{i}\frac{d}{dx}$ with a complex-valued potential $V \in H_0^w$ of period 1, considered as an unbounded operator on $L^2_{\mathbb{C}}[0,2]$, with periodic boundary conditions. For $V = 0$, the spectrum is discrete: $0, \pi^2, \pi^2, (2\pi)^2, (2\pi)^2, \dots$; i.e., the eigenvalues $k^2\pi^2$ are double for $k \geq 1$ and the eigenvalues $(n+1)^2\pi^2$ and $n^2\pi^2$ are $(2n+1)\pi^2$ apart. Further, for $n \geq 1$, $e^{in\pi x}, e^{-in\pi x}$ is a basis of the eigenspace corresponding to the eigenvalue $n^2\pi^2$. Viewing the potential V as a perturbation of D^2, it follows that for n sufficiently large, L has a pair of eigenvalues near $n^2\pi^2$, isolated from the remaining spectrum of L. Our aim is to obtain an estimate for the distance between the two eigenvalues and to compare eigenfunctions and eigenvalues with the corresponding ones for $V = 0$. Notice, however, that L might have double eigenvalues of geometric multiplicity 1 as V is complex-valued.

The Fourier series decomposition leads to an isometric isomorphism $\mathcal{F}$ between $L^2_{\mathbb{C}}[0,2]$ and $\ell^2(\mathbb{Z})$ with $\mathcal{F}(e^{i\pi kx}) = e_k$, $(e_k)_{k\in\mathbb{Z}}$ being the standard basis in $\ell^2(\mathbb{Z})$. Decompose $\hat{L} = \mathcal{F}L\mathcal{F}^{-1}$ with respect to the orthogonal sum $\ell^2(\mathbb{Z}) = \mathbb{C}e_{-n} \oplus \mathbb{C}e_n \oplus \ell^2(\mathbb{Z}\backslash\{\pm n\})$. To express $\hat{L}$, introduce the involution operator $J : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$,

$$(Ja)(k) := a(-k)\ (k \in \mathbb{Z})$$

and the shift operator $\mathcal{S} : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$,

$$(\mathcal{S}a)(k) := a(k+1)\ (k \in \mathbb{Z}).$$

$\mathcal{S}^n = \mathcal{S} \circ \cdots \circ \mathcal{S}$ denotes the n*th iterate of $\mathcal{S}$. For any subset $K \subset \mathbb{Z}$, the restriction of $\mathcal{S}$ on $\ell^2(K)$ with values in $\ell^2(\mathcal{S}(K))$ is denoted by $\mathcal{S}$ as well. This leads to the block decomposition of $\hat{L} - \lambda, \lambda = n^2\pi^2 + z$,

$$\hat{L} - (n^2\pi^2 + z) = \begin{pmatrix} -z & \hat{V}(-2n) & (\mathcal{S}^n J\hat{V})^t_{\mathbb{Z}(n)} \\ \hat{V}(2n) & -z & (\mathcal{S}^{-n} J\hat{V})^t_{\mathbb{Z}(n)} \\ (\mathcal{S}^n \hat{V})_{\mathbb{Z}(n)} & (\mathcal{S}^{-n}\hat{V})_{\mathbb{Z}(n)} & A_n - z \end{pmatrix}, \tag{2.1}$$

where $\mathbb{Z}(n) := \mathbb{Z}\backslash\{\pm n\}$, the superscript t denotes the transpose, and $A_n : \ell^2(\mathbb{Z}\backslash\{\pm n\}) \to \ell^2(\mathbb{Z}\backslash\{\pm n\})$ is the linear operator with matrix representation

$$A_n(j,k) = \pi^2(k^2 - n^2)\delta_{jk} + \hat{V}(j-k), \quad (j,k \in \mathbb{Z}(n)).$$

The (possibly) complex number $\lambda = n^2\pi^2 + z$ is a periodic eigenvalue of L if there exists a 2-periodic function $f \in H^2_{\mathbb{C}}([0,2])$ such that $(L - \lambda)f = 0$. With

$$x^f := \hat{f}(-n), \quad y^f := \hat{f}(n), \quad F := (\hat{f}(k))_{\mathbb{Z}(n)},$$

the equation $(L - \lambda)f = 0$, or its equivalent $(\hat{L} - \lambda)\hat{f} = 0$, leads to the following homogeneous system of equations:

$$-zx^f + \hat{V}(-2n)y^f + [\mathcal{S}^n J\hat{V}, F]_{\mathbb{Z}(n)} = 0, \tag{2.2}$$

$$\hat{V}(2n)x^f - zy^f + [\mathcal{S}^{-n} J\hat{V}, F]_{\mathbb{Z}(n)} = 0, \tag{2.3}$$

$$(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)}x^f + (\mathcal{S}^{-n}\hat{V})_{\mathbb{Z}(n)}y^f + (A_n - z)F = 0, \tag{2.4}$$

where $[a,b]_K = \sum_{k\in K} a(k)b(k)$ (no complex conjugation). Equation (2.4) will be referred to as *the external equation.* The system of equations (2.2)–(2.4) is analyzed as follows. First we solve the external equation (2.4) for F, regarding x^f, y^f, and z as parameters. The solution F of (2.4) is then substituted into the equations (2.2)–(2.3). This leads to a linear homogeneous system of two equations for the unknowns x^f, y^f with parameter z. The determinant of this system vanishes iff $\lambda = n^2\pi^2 + z$ is an eigenvalue of L. In section 3.3 we will also consider the inhomogneous version of the system (2.2)–(2.4) in order to obtain, among other results, an orthonormal basis of the root space of a double eigenvalue of L of geometric multiplicity 1.

2.2. Analysis of the external equation. To analyze the operator $(A_n - z) : \ell^2(\mathbb{Z}(n)) \to \ell^2(\mathbb{Z}(n))$, we write $A_n = D_n + B_n$, where D_n is the diagonal part of A_n (recall that $\hat{V}(0) = 0$),

$$D_n(k,j) := \pi^2(k^2 - n^2)\delta_{kj}, \quad (k,j \in \mathbb{Z}(n) = \mathbb{Z}\backslash\{\pm n\}). \tag{2.5}$$

Notice that D_n is invertible and that B_n has matrix elements

$$B_n(k,j) = \hat{V}(k-j), \quad (k,j \in \mathbb{Z}(n)).$$

Write

$$A_n - z = D_n - (z - B_n) = (Id - T_n)D_n; \quad T_n := (z - B_n)D_n^{-1}, \tag{2.6}$$

where T_n is an operator on $\ell^2(\mathbb{Z}(n))$ with matrix elements

$$(z\delta_{kj} - \hat{V}(k-j))\frac{1}{\pi^2(j - n^2)}.$$

Further, denote by $\|V\|$ the norm of V in $L^2_{\mathbb{C}}([0,2]), \|V\| = (2\sum_k |\hat{V}(k)|^2)^{1/2}$.

Lemma 2.1. (i) *For $n \geq 1$,*

$$\|D_n^{-1}\| \;\leq\; \frac{1}{\pi^2}\frac{1}{n}, \tag{2.7}$$

$$\|T_n\| \;\leq\; \frac{1}{3n}(|z| + \|V\|); \tag{2.8}$$

(ii) *for $n \geq 1$ and $z \in \mathbb{C}$ with $|z| + \|V\| \;\leq\; n$,*

$$\|(A_n - z)^{-1}\| \;\leq\; \frac{2}{\pi^2}\frac{1}{n}. \tag{2.9}$$

Proof. (i) (2.7) follows from (2.5). Concerning (2.8) we prove $\|T_n\|_{HS} \leq \frac{1}{3n}(|z| + \|V\|)$ with $\|T_n\|_{HS}$ denoting the Hilbert–Schmidt norm of T (which leads to a stronger version of (2.8) as $\|T_n\| \leq \|T_n\|_{HS}$):

$$\begin{aligned} \|T_n\|^2_{HS} &= \sum_{j,k\in\mathbb{Z}(n)} \frac{|z\delta_{kj} - \hat{V}(k-j)|^2}{|\pi^2(k^2-n^2)|^2} \\ &\leq \sum_{k\in\mathbb{Z}(n)} \frac{1}{\pi^4}\frac{2|z|^2 + 2\|\hat{V}\|^2}{(k-n)^2(k+n)^2}. \end{aligned}$$

As

$$\sum_{k\neq\pm n}\frac{1}{(k-n)^2(k+n)^2}=\frac{1}{6}\left(\frac{\pi}{n}\right)^2-\frac{3}{8}\frac{1}{n^4}, \tag{2.10}$$

we conclude that $\|T_n\|_{HS}\leq\frac{1}{3n}(|z|+\|V\|)$.

(ii) If $\|T_n\|\leq\frac{1}{2}, (A_n-z)$ is invertible (cf. (2.6)) and

$$\|(A_n-z)^{-1}\| \;\leq\; 2\|D_n^{-1}\| \;\leq\; \frac{2}{\pi^2}\frac{1}{n}.$$

In view of (2.8), $\|T_n\|\leq 1/2$ for $|z|+\parallel V\parallel\leq n$. □

As an immediate consequence of Lemma 2.1, one obtains the following proposition.

PROPOSITION 2.2. *Let $n\geq 1$ and $z\in\mathbb{C}$ satisfy $|z|+\parallel V\parallel\leq n$. Then, for any choice of x^f, y^f in $\mathbb{C}$, (2.4) has a unique solution F*

$$F=(z-A_n)^{-1}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)}x^f+(z-A_n)^{-1}(\mathcal{S}^{-n}\hat{V})_{\mathbb{Z}(n)}y^f. \tag{2.11}$$

Substituting the solution F, given by (2.11), into (2.2)–(2.3), one gets

$$\begin{pmatrix} -z+\alpha(-n,z) & \hat{V}(-2n)+\beta(-n,z)\\ \hat{V}(2n)+\beta(n,z) & -z+\alpha(n,z)\end{pmatrix}\begin{pmatrix} x^f\\ y^f\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}, \tag{2.12}$$

where, for $n\in\mathbb{Z}\backslash\{0\}$, satisfying $|z|+\|V\|\leq n$, we define, with $A_n:=A_{|n|}$,

$$\alpha(n,z):=[\mathcal{S}^{-n}J\hat{V},(z-A_n)^{-1}(\mathcal{S}^{-n}\hat{V})_{\mathbb{Z}(n)}]_{\mathbb{Z}(n)}, \tag{2.13}$$

$$\beta(n,z):=[\mathcal{S}^{-n}J\hat{V},(z-A_n)^{-1}(\mathcal{S}^{n}\hat{V})_{\mathbb{Z}(n)}]_{\mathbb{Z}(n)}. \tag{2.14}$$

In the following sections, the coefficients $\alpha(n,z)$ and $\beta(n,z)$ will be analyzed. Often, we will write $[\cdot,\cdot]$, for $[\cdot,\cdot]_{\mathbb{Z}(n)}$.

2.3. Identity for $\boldsymbol{\alpha(n,z)}$. Throughout this and the following section, we assume that $n\geq 1$ and $z\in\mathbb{C}$ are such that $|z|+\|V\|\leq n$. To simplify notation we drop the subindex n in A_n, B_n, D_n, and T_n with the understanding that n is fixed in this subsection. Denote by $(z-A)^t$ the transpose of $z-A, (z-A)^t(j,k):=(z-A)(k,j)(\forall k,j\in\mathbb{Z}(n))$.

LEMMA 2.3. (i) $J\mathcal{S}^n=\mathcal{S}^{-n}J$;

(ii) $(z-A)=J(z-A)^tJ$.

Proof. Part (i) is verified in a straightforward way. Regarding (ii), it is to prove that for $k,j\in\mathbb{Z}(n)$,

$$(z-A)(k,j)=(z-A)(-j,-k). \tag{2.15}$$

The identity (2.15) follows from the definition

$$(z-A)(k,j)=z\delta_{kj}-\pi^2(k^2-n^2)\delta_{kj}-\hat{V}(k-j)$$

and the identities

$$\delta_{kj}=\delta_{(-j)(-k)};\quad \hat{V}(k-j)=\hat{V}(-j-(-k)). \qquad □$$

We obtain the following identity for $\alpha(n,z)$.

LEMMA 2.4. $\alpha(n,z) = \alpha(-n,z)$.

Proof. With $(\mathcal{S}^{-n}\hat{V})_{\mathbb{Z}(n)} = \mathcal{S}^{-n}\hat{V}_{\mathbb{Z}\setminus\{0,2n\}}$ and $(z-A)^{-1} = (J(z-A)^tJ)^{-1} = J((z-A)^{-1})^tJ$ (Lemma 2.3) it follows that

$$\begin{aligned}\alpha(n,z) &= [\mathcal{S}^{-n}J\hat{V}, (J(z-A)^tJ)^{-1}\mathcal{S}^{-n}\hat{V}_{\mathbb{Z}\setminus\{0,2n\}}]_{\mathbb{Z}(n)} \\ &= [J\mathcal{S}^{n}\hat{V}, J((z-A)^{-1})^tJ\mathcal{S}^{-n}\hat{V}_{\mathbb{Z}\setminus\{0,2n\}}]_{\mathbb{Z}(n)} \\ &= [(z-A)^{-1}\mathcal{S}^{n}\hat{V}, J\mathcal{S}^{-n}\hat{V}_{\mathbb{Z}\setminus\{0,2n\}}]_{\mathbb{Z}(n)} \\ &= [\mathcal{S}^{n}J\hat{V}, (z-A)^{-1}\mathcal{S}^{n}\hat{V}]_{\mathbb{Z}(n)} \\ &= \alpha(-n,z). \qquad \square\end{aligned}$$

As a consequence of Lemma 2.4, the vanishing of the determinant of the 2 × 2 matrix in (2.12) leads to the following equation for z:

$$(z-\alpha(n,z))^2 - \Big(\hat{V}(2n)+\beta(n,z)\Big)\Big(\hat{V}(-2n)+\beta(-n,z)\Big) = 0. \tag{2.16}$$

Equation (2.16) is solved in two steps: for ζ given, we first solve the following equation, referred to as the z-equation, for z:

$$z = \alpha(n,z)+\zeta. \tag{2.17}$$

Substituting the solution $z = z(\zeta)$ of (2.17) into (2.16), we obtain the following equation for ζ, referred to as the ζ-equation:

$$\zeta^2 - \Big(\hat{V}(2n)+\beta(n,z(\zeta))\Big)\Big(\hat{V}(-2n)+\beta(-n,z(\zeta))\Big) = 0. \tag{2.18}$$

In the next four sections, (2.17) and (2.18) will be analyzed.

2.4. Estimates of $\alpha(n,z)$ and the z-equation (2.17).

In this section we solve (2.17), using the contractive mapping principle. For this purpose we need the following lemma.

LEMMA 2.5. *For $n \geq 1$ and $z \in \mathbb{C}$ satisfying $|z| + \|V\| \leq n$,*

(i) $|\alpha(n,z)| \leq \|V\|^2/3n$;

(ii) $|\frac{d}{dz}\alpha(n,z)| \leq \|V\|^2/9n^2$.

Proof. (i) By the definition (2.13) and Lemma 2.1,

$$|\alpha(n,z)| \leq \|V\|\|(z-A)^{-1}\|\,\|V\| \leq \frac{2}{\pi^2}\frac{1}{n}\|V\|^2.$$

(ii) Notice that

$$\frac{d}{dz}\alpha(n,z) = [\mathcal{S}^{-n}J\hat{V}, -(z-A)^{-2}\mathcal{S}^{-n}\hat{V}]_{\mathbb{Z}(n)},$$

and therefore,

$$\left|\frac{d}{dz}\alpha(n,z)\right| \leq \|V\|\|(z-A)^{-1}\|^2\|V\| \leq \frac{4}{\pi^4}\frac{1}{n^2}\|V\|^2. \qquad \square$$

Denote by $\mathcal{D}_M \equiv \mathcal{D}_M(0)$ the disc $\{z \in \mathbb{C} | |z| < M\}$ and denote by $\overline{\mathcal{D}_M}$ its closure.

Proposition 2.6. *Let $V \in L^2_0$. Then for any $M > 0$ and $n \geq 1$ satisfying $n \geq \|V\| + M$, and for any $\zeta \in \mathcal{D}_{M/2}$, the equation*

$$z = \zeta + \alpha(n, z) \tag{2.19}$$

has a unique solution $z_n = z_n(\zeta)$ in $\mathcal{D}_M$. The solution $z_n(\zeta)$ depends analytically on $\zeta \in \mathcal{D}_{M/2}$.

Proof. For $z \in \overline{\mathcal{D}_M}$,

$$|z| + \|V\| \leq M + \|V\| \leq n,$$

and thus, by Lemma 2.5, $|\alpha(n, z)| \leq M/3$. It follows that for $\zeta \in \mathcal{D}_{M/2}, z \in \overline{\mathcal{D}_M}$

$$|\zeta| + |\alpha(n, z)| \leq M/2 + M/3 < M.$$

Thus, for $\zeta \in \mathcal{D}_{M/2}$, $g(z) := \zeta + \alpha(n, z)$ defines a map on $\overline{\mathcal{D}_M}$ into $\overline{\mathcal{D}_M}$. Furthermore, g is a contraction, as for any $z_1, z_2 \in \mathcal{D}_M$

$$|g(z_1) - g(z_2)| < \frac{1}{9}|z_1 - z_2|,$$

where we used that by Lemma 2.5

$$\sup_{|z| \leq M} \left| \frac{d}{dz}\alpha(n, z) \right| \leq \frac{1}{9}\frac{1}{n^2}\|V\|^2 \leq \frac{1}{9}.$$

Hence there exists a fixed point $z = z(\zeta)$ of g with $|z| \leq M$, z with

$$\frac{dz}{d\zeta} = \left(1 - \frac{d\alpha}{dz}(n, z(\zeta))\right)^{-1} \quad \forall |\zeta| < M/2. \qquad \square$$

In a next step, we analyze (2.18). To obtain estimates for the coefficient $\beta(n, z)$ we need to establish bounds for the norm of the operator $T = (z - B)D^{-1}$ introduced in section 2.2, viewed as an operator on a weighted ℓ^2-space.

2.5. Estimates of norms of T_n. Recall that, with n arbitrary and fixed, $T \equiv T_n = (z - B)D^{-1} : \ell^2(\mathbb{Z}(n)) \to \ell^2(\mathbb{Z}(n))$ is a bounded operator (cf. (2.6)). If $V \in H^w_0, T$ can also be viewed as an element in $\mathcal{L}(\ell^2_{\mathcal{S}^{\pm n}w}(\mathbb{Z}(n)))$, where $\mathcal{S}^{\pm n}w$ is the shifted weight

$$(\mathcal{S}^{\pm n}w)(j) := w(\pm n + j). \tag{2.20}$$

Denote by $W_\pm : \ell^2_{\mathcal{S}^{\pm n}w}(\mathbb{Z}(n)) \to \ell^2(\mathbb{Z}(n))$ the diagonal operator given by

$$W_\pm(k, j) = w(k \pm n)\delta_{kj}.$$

Notice that $W_\pm : \ell^2_{\mathcal{S}^{\pm n}w}(\mathbb{Z}(n)) \to \ell^2(\mathbb{Z}(n))$ is an isometry. Therefore,

$$\|T_\pm\|_{\mathcal{L}(\ell^2)} = \|T_n\|_{\mathcal{L}(\ell^2_{\mathcal{S}^{\pm n}w})}, \tag{2.21}$$

where $T_\pm := W_\pm T_n W_\pm^{-1} : \ell^2(\mathbb{Z}(n)) \to \ell^2(\mathbb{Z}(n))$.

Lemma 2.7. *For $n \geq 1$*

$$\|T_n\|_{\mathcal{L}(\ell^2_{\mathcal{S}^{\pm n}w})} \leq \frac{|z| + \|V\|_w}{3n}.$$

Proof. In view of (2.21) it suffices to estimate the Hilbert–Schmidt norm of $T_\pm$ in $\mathcal{L}(\ell^2)$. As w is submultiplicative,

$$\frac{(\mathcal{S}^{\pm n}w)(j)}{(\mathcal{S}^{\pm n}w)(k)} \le w(j-k).$$

In view of (2.21), (2.6), and $\hat{V}(0) = 0$,

$$\begin{aligned}\|T_\pm\|_{HS}^2 &= \sum_{j,k\neq\pm n} \left|\frac{\mathcal{S}^{\pm n}w(j)}{\mathcal{S}^{\pm n}w(k)}\right|^2 |\hat{V}(j-k)|^2 \frac{1}{\pi^4|k^2-n^2|^2} \\ &\quad + \sum_{k\neq\pm n} |z|^2 \frac{1}{\pi^4|k^2-n^2|^2} \\ &\le (\|V\|_w^2 + |z|^2) \sum_{k\neq\pm n} \frac{1}{\pi^4} \frac{1}{(k-n)^2(k+n)^2} \\ &\le \frac{|z|^2+\|V\|_w^2}{9}\frac{1}{n^2},\end{aligned}$$

where in the last inequality, we again use (2.10). This estimate leads to $\|T_\pm\|_{HS} \le \frac{|z|+\|V\|_n}{3n}$. □

As an immediate consequence of Lemma 2.7 one obtains the following corollary.

COROLLARY 2.8. *For $n \ge 1$, and $z \in \mathbb{C}$ with $|z|+\|V\|_w \le n$, $Id-T_n$ is invertible and*

$$\|(Id-T_n)^{-1}\|_{\mathcal{L}(\ell^2_{\mathcal{S}^{\pm n}w})} \le 2.$$

Corollary 2.8 can be used to obtain an estimate of the solution F of the external equation established in Proposition 2.2. According to (2.11),

$$F = x^f F_+ + y^f F_-,$$

where

$$F_\pm = (z-A_n)^{-1}(\mathcal{S}^{\pm n}\hat{V})_{\mathbb{Z}(n)}. \tag{2.22}$$

COROLLARY 2.9. *For $n \ge 1$ and $z \in \mathbb{C}$ with $|z| + \|V\|_w \le n$,*

$$\|F_\pm\|_{\ell^2_{\mathcal{S}^{\pm n}w}} \le \frac{2}{\pi^2 n}\|V\|_w.$$

Proof. By (2.6), $(z-A_n)^{-1} = -D_n^{-1}(Id-T_n)^{-1}$, and by Corollary 2.8

$$\|(Id-T_n)^{-1}\|_{\mathcal{L}(\ell^2_{\mathcal{S}^{\pm n}w})} \le 2.$$

As D_n is the diagonal operator on $\ell^2(\mathbb{Z}(n))$ with coefficients $D_n(k,j) = \pi^2(k^2-n^2)\delta_{kj}$, we have

$$\|D_n^{-1}\|_{\mathcal{L}(\ell^2_{\mathcal{S}^{\pm n}w})} \le \frac{1}{\pi^2 n}.$$

Combining these estimates yields the claimed estimate. □

2.6. Estimate for $\boldsymbol{\beta(n,z)}$. Substitute, for $z \in \mathbb{C}$ satisfying $|z| + \|V\| \leq n$,

$$(z - A)^{-1} = -D^{-1} - D^{-1}T(Id - T)^{-1}$$

into the expression for $\beta(n,z)$ to obtain

$$\begin{aligned} \beta(n,z) &= \beta_1(n) + \beta_2(n,z) \\ &= \beta_1(n) + \beta_2(n,0) + z\int_0^1 \left(\frac{d}{dz}\beta\right)(n,tz)dt, \end{aligned} \tag{2.23}$$

where

$$\beta_1(n) := -[\mathcal{S}^{-n}J\hat{V}, D^{-1}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)}]_{\mathbb{Z}(n)}, \tag{2.24}$$

$$\beta_2(n,z) := -[\mathcal{S}^{-n}J\hat{V}, D^{-1}T(Id - T)^{-1}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)}]_{\mathbb{Z}(n)}. \tag{2.25}$$

The term $\beta_1(n)$ is independent of z and

$$\beta_1(n) = \frac{1}{\pi^2}\left(\frac{\hat{V}}{k} * \frac{\hat{V}}{k}\right)(2n), \tag{2.26}$$

or

$$\beta_1(n) = \frac{1}{\pi^2}\sum_{k \neq \pm n} \frac{\hat{V}(n-k)}{n-k}\frac{\hat{V}(n+k)}{n+k}, \tag{2.27}$$

where we use that $-D(k,j) = -\pi^2(k^2 - n^2)\delta_{kj} = \pi^2(n-k)(n+k)\delta_{kj}$. In the subsequent lemmas, $\beta_1(n), \beta_2(n,0)$, and $\frac{d}{dz}\beta(n,z)$ are estimated separately. Given the weight w and $\alpha > 1/2$, introduce a new weight $(w_\alpha(k))_{k\in\mathbb{Z}}$,

$$w_\alpha(k) := \left(1 + |\frac{k}{2}|\right)^\alpha w(k).$$

Notice that w_α is again symmetric and submultiplicative.

LEMMA 2.10. $(\sum_{n\in\mathbb{Z}} w_1(2n)^2\beta_1(n)^2)^{1/2} \leq \|V\|_w^2$.

Proof. By Lemmas A.1 and A.2 (in particular, Lemma A.2 for $\alpha = 1$) and (2.26),

$$\left(\sum_{n\in\mathbb{Z}} w_1(2n)^2\beta_1(n)^2\right)^{1/2} \leq \frac{1}{\pi^2}\left\|\frac{\hat{V}}{k} * \frac{\hat{V}}{k}\right\|_{w_1} \leq \frac{6}{\pi^2}\left\|\frac{\hat{V}}{k}\right\|_{w_1}^2 \leq \|\hat{V}\|_w^2. \qquad \square$$

LEMMA 2.11. *For* $|n| \geq n_w := M + \|V\|_w$,

$$(1 + |n|)w(2n)\sup_{|z|\leq M}|\beta_2(n,z)| \leq \frac{1}{3}\|V\|_w^2.$$

Proof. By (2.25), for any $|z| \leq M$ and $|n| \geq n_w$,

$$\begin{aligned} |\beta_2(n,z)| &= \frac{1}{\pi^2}\left|\left(\frac{\hat{V}}{k} * \frac{\mathcal{S}^{-n}T(Id-T)^{-1}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)}}{k}\right)(2n)\right| \\ &\leq \frac{1}{\pi^2}\sum_{k\neq\pm n}\frac{|\hat{V}(n-k)|}{|n-k|}\frac{|a_{(n)}(n+k)|}{|n+k|}, \end{aligned} \tag{2.28}$$

where

$$a_{(n)}(k) := \mathcal{S}^{-n}T(Id-T)^{-1}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)}(k), \quad (k \in \mathbb{Z}\backslash\{-2n, 0\}).$$

Using that $\|T\|_{\mathcal{L}(\ell^2_{\mathcal{S}^n w})} \leq \frac{M+\|V\|_w}{3|n|} = \frac{n_w}{3|n|}$ (Lemma 2.7) and $\|(Id-T)^{-1}\|_{\mathcal{L}(\ell^2_{\mathcal{S}^n_w})} \leq 2$ (Corollary 2.8), we conclude that

$$\|a_{(n)}\|_w \leq \frac{n_w}{3|n|}2\|V\|_w \leq \frac{2}{3}\|V\|_w. \tag{2.29}$$

As w_1 is submultiplicative, we then obtain from (2.28)

$$\begin{aligned}&(1+|n|)w(2n)|\beta_2(n,z)|\\ &\leq \frac{1}{\pi^2}\sum_{k\neq\pm n} 2w(n-k)|\hat{V}(n-k)|2w(n+k)|a_{(n)}(n+k)|\\ &\leq \frac{4}{\pi^2}\|V\|_w\|a_{(n)}\|_w \leq \frac{1}{3}\|V\|^2_w. \qquad \square\end{aligned}$$

Lemma 2.12.

$$\left(\sum_{|n|\geq n_w}(1+|n|)^4 w(2n)^2|\beta_2(n,0)|^2\right)^{1/2} \leq (1+n_w)\|V\|^3_w.$$

Proof. By (2.25),

$$\beta_2(n,0) = [\mathcal{S}^{-n}J\hat{V}, D^{-1}BD^{-1}(Id-T_{z=0})^{-1}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)}]_{\mathbb{Z}(n)}.$$

Write $(Id-T_{z=0})^{-1} = Id + T_{z=0}(Id-T_{z=0})^{-1}$ to obtain $\beta_2(n,0) = \beta_3(n) + \beta_4(n)$ with

$$\begin{aligned}\beta_3(n) :=&[\mathcal{S}^{-n}J\hat{V}, D^{-1}BD^{-1}\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)}]_{\mathbb{Z}(n)},\\ \beta_4(n) :=&[\mathcal{S}^{-n}J\hat{V}, D^{-1}BD^{-1}\mathcal{S}^n a_{(n)}]_{\mathbb{Z}(n)},\end{aligned}$$

and

$$a \equiv a_{(n)} := \mathcal{S}^{-n}T_{z=0}(Id-T_{z=0})^{-1}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)}.$$

Let us first treat β_3. As w is submultiplicative,

$$w(2n) \leq w(n-k)w(n+k) \leq w(n-k)w(n+j)w(k-j),$$

one obtains, for $|n| \geq n_w$,

$$\begin{aligned}&(1+|n|)^2 w(2n)\beta_3(n)\\ &\leq \frac{1}{\pi^4}\sum_{\substack{k\neq\pm n\\ j\neq\pm n}}(1+|n|)^2\frac{w(n-k)|\hat{V}(n-k)|}{|n-k||n+k|}w(k-j)|\hat{V}(k-j)|\frac{w(n+j)|\hat{V}(n+j)|}{|n-j||n+j|}\\ &= \frac{1}{\pi^4}(R_1+R_2+R_3+R_4),\end{aligned} \tag{2.30}$$

where R_1, R_2, R_3, R_4 denote the partial sums corresponding to the index sets I_1, I_2, I_3, $I_4 \subseteq \{(k,j) \in \mathbb{Z}^2 | k \neq \pm n, j \neq \pm n\}$ defined as follows:

$$\begin{aligned} I_1 &:= \{|k-n| > |n|; \quad |j-n| > |n| \quad | \quad k, j \neq \pm n\},\\ I_2 &:= \{|k-n| > |n|; \quad |j-n| < |n| \quad | \quad k, j \neq \pm n\},\\ I_3 &:= \{|k-n| < |n|; \quad |j-n| > |n| \quad | \quad k, j \neq \pm n\},\\ I_4 &:= \{|k-n| < |n|; \quad |j-n| < |n| \quad | \quad k, j \neq \pm n\}. \end{aligned}$$

Let us first estimate $R_2(n)$. For k, j with $|k-n| > |n|, |j-n| < |n|$, one has $1 + |n| \leq |k-n|$ and $1 + |n| \leq |j+n|$, and thus,

$$\frac{(1+|n|)^2}{|k-n||j+n|} \leq \frac{|k-n|}{|k-n|}\frac{|j+n|}{|j+n|} \leq 1.$$

Therefore, $R_2(n)$ is bounded by

$$\sum_{(k,j)\in I_2} \frac{w(n-k)|\hat{V}(n-k)|}{|n+k|} w(k-j)|\hat{V}(k-j)| \frac{w(n+j)|\hat{V}(n+j)|}{|n-j|}. \tag{2.31}$$

Let $\xi(j) := w(j)|\hat{V}(j)|$. Then

$$\sum_{\substack{|j-n|<|n|\\ j\neq n}} \xi(k-j)\frac{\xi(n+j)}{|n-j|} = \sum_{\substack{|\ell-2n|<|n|\\ 2n-\ell\neq 0}} \xi(k+n-\ell)\frac{\xi(\ell)}{|2n-\ell|} \leq \rho. \tag{2.32}$$

Hence, for $h \in \ell_2(\mathbb{Z})$,

$$\begin{aligned} &\sum_{|n|\geq n_w} |h(n)| R_2(n) \leq \sum_{n,k,\ell} |h(n)| \frac{\xi(n+k)}{|k+n|} \frac{\xi(k+n-\ell)\xi(\ell)}{|2n-\ell|}\\ &= \sum_{n,k,\ell} \frac{h(n)\xi(k+n-\ell)}{2n-\ell} \frac{\xi(n-k)\xi(\ell)}{|k+n|}\\ &\leq \left(\sum_{n,k,\ell} \frac{|h(n)|^2 \xi(k+n-\ell)^2}{|2n-\ell|^2}\right)^{1/2} \left(\sum_{n,k,\ell} \frac{\xi(n-k)^2\xi(\ell)^2}{|k+n|^2}\right)^{1/2}\\ &\leq \left(\sum_{n,j,\ell} \frac{|h(n)|^2 \xi(j)^2}{|2n-\ell|^2}\right)^{1/2} \left(\sum_{j,\ell,n} \frac{\xi(j)^2\xi(\ell)^2}{|2n-j|^2}\right)^{1/2}\\ &\leq \cdot 2\|h\| \ \|\xi\| \cdot 2\|\xi\|^2 = 4\|h\| \ \|\xi\|^3. \end{aligned}$$

Thus we have proved that

$$\left(\sum_{|n|\geq n_w} R_2(n)^2\right)^{1/2} \leq 4\|\xi\|^3 \leq 4\|V\|_w^3.$$

Using the convolution estimate $\|U * V\|_{\ell^2} \leq \|U\|_{\ell^2}\|V\|_{\ell^1}$ one obtains, for $j = 1$, 3, and 4,

$$\left(\sum_{|n|\geq n_w} R_j(n)^2\right)^{1/2} \leq 4\|V\|_w^3.$$

Hence,

$$\begin{aligned}&\left(\sum_{|n|\geq n_w}(1+|n|)^4 w(2n)^2|\beta_3(n)|^2\right)^{1/2}\\ &\leq \frac{1}{\pi^4}4\cdot 4\cdot \|V\|_w^3 \leq \|V\|_w^3.\end{aligned}$$

To estimate $\beta_4(n)$ we proceed similarly. By definition,

$$\beta_4(n)=\sum_{k\neq\pm n}\hat{V}(n\leq k)\sum_{j\neq\pm n}\frac{1}{\pi^2(k^2-n^2)}\hat{V}(k-j)\frac{1}{\pi^2(j^2(j^2-n^2)}a(n+j),$$

whence

$$\begin{aligned}&(1+|n|)^2 w(2n)|\beta_4(n)|\\ &\leq \frac{1}{\pi^4}\sum_{\substack{k\neq\pm n\\ j\neq\pm n}}(1+|n|)^2\frac{w(n-k)\hat{V}(n-k)}{|n-k|\;|n+k|}w(k-j)\hat{V}(k-j)\frac{w(n+j)|a(n+j)|}{|n-j|\;|n+j|}\\ &=\frac{1}{\pi^4}(Q_1+Q_2+Q_3+Q_4),\end{aligned}$$

where Q_1, Q_2, Q_3, Q_4 denote the partial sums corresponding to the index sets I_1, I_2, I_3, I_4 defined above. Each of the four terms $Q_i = Q_i(n)(1 \leq i \leq 4)$ is estimated in the same way, so we concentrate only on one of them, say, Q_2. Similarly as in (2.31) we obtain

$$\sum_{(k,j)\in I_2}\frac{w(n-k)|\hat{V}(n-k)|}{|n+k|}w(k-j)|\hat{V}(k-j)|w(n+j)\frac{|a(n+j)|}{|n-j|}.$$

Let $\eta(j)\equiv\eta_{(n)}(j) := w(j)|a(j)|$ and $\xi(j) := w(j)|\hat{V}(j)|$. Then

$$\sum_{\substack{|j-n|<|n|\\ j\neq n}}\xi(k-j)\frac{\eta(n+j)}{|n-j|}=\sum_{\substack{|\ell-2n|<|n|\\ 2n-\ell\neq 0}}\xi(k+n-\ell)\frac{\eta(\ell)}{|2n-\ell|}\leq\delta_{(n)}(n+k),$$

where

$$\delta_{(n)}(k+n) := \sum_{\ell\neq 2n,0}\xi(k+n-\ell)\frac{\eta_{(n)}(\ell)}{|2n-\ell|}.$$

Using the convolution estimate $\|U * V\|_{\ell^2} \leq \|U\|_{\ell^2}\|V\|_{\ell^1}$ one concludes that (with $2\cdot\sum_{j\geq 1}\frac{1}{j^2} = 2\cdot\frac{\pi^2}{6} < 4$)

$$\|\delta_{(n)}\|_{\ell^2}\leq\|\eta_{(n)}\|_{\ell^2}\|\xi\|_{\ell^2}\left(\sum_{j\neq 0}\frac{1}{j^2}\right)^{1/2}\leq 2\|\eta_{(n)}\|_{\ell^2}\|\xi\|_{\ell^2}.$$

As $\|T_{Z=0}\|_{\mathcal{L}(\ell^2_{\mathcal{S}^n w})} \leq \frac{n_w}{3|n|}$ (Lemma 2.7) and $\|(Id{-}I_{Z=0})^{-1}\|_{\mathcal{L}(\ell^2_{\mathcal{S}^n w})} \leq 2$ (Corollary 2.8), one has

$$\|\eta_{(n)}\|_{\ell^2}\leq\frac{nw}{3|n|}1\|V\|_w,\quad(\forall|n|\geq n_w).$$

Hence,

$$\|\delta_{(n)}\|_{\ell^2} \le \frac{4n_w}{3|n|}\|V\|_w^2, \quad (\forall |n| \ge n_w).$$

This leads to

$$Q_2(n) \le \sum_{\substack{|k-n|>|n| \\ k\neq -n}} \xi(n-k)\frac{\delta_{(n)}(k+n)}{|k+n|} \le \|\xi\|_{\ell^2}\frac{4n_w}{3|n|}\|V\|_w^2 \cdot 2,$$

and hence,

$$\begin{aligned}\left(\sum_{|n|\ge w_w} Q_2(n)^2\right)^{1/2} &\le \frac{8n_w}{3}\|V\|_w^3\left(\sum_{|n|\ge n_w}\frac{1}{n^2}\right)^{1/2} \\ &\le \frac{16n_w}{3}\|V\|_w^3.\end{aligned}$$

Similar estimates hold for Q_1, Q_3, and Q_4, and thus,

$$\begin{aligned}&\left(\sum_{|n|\ge n_w}(1+|n|)^4 w(2n)^2|\beta_4(n)|^2\right)^{1/2} \\ &\le 4\cdot\frac{1}{\pi^4}\cdot\frac{16n_w}{3}\|V\|_w^3.\end{aligned}$$

Combined with the estimate for $\beta_3(n)$, this leads to the claimed statement. ∎

LEMMA 2.13.

$$\left(\sum_{|n|\ge n_w}(1+|n|)^4 w(2n)^2 \sup_{|z|\le M}\left|\frac{d}{dz}\beta(n,z)\right|^2\right)^{1/2} \le 2(1+n_w)^{1/2}\|V\|_w^2.$$

Proof. Let $\eta(n,z) := \frac{d}{dz}\beta(n,z)$ and notice that

$$\eta(n,z) = -[\mathcal{S}^{-n}J\hat{V}, (z-A)^{-2}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)}]_{\mathbb{Z}(n)}.$$

Recall that $(z-A)^{-1} = -D^{-1}(Id-T)^{-1} = -D^{-1} - D^{-1}T(Id-T)^{-1}$ and thus

$$\begin{aligned}(z-A)^{-2} &= \left(-D^{-1} - D^{-1}T(Id-T)^{-1}\right)(z-A)^{-1} \\ &= D^{-2} + D^{-2}T(Id-T)^{-1} - D^{-1}T(Id-T)^{-1})(z-A)^{-1} \\ &= D^{-2} + D^{-2}T(Id-T)^{-1} + D^{-1}T(Id-T)^{-1}D^{-1}(Id-T)^{-1}.\end{aligned}$$

This is used to write $\eta(n,z)$ as a sum,

$$\eta(n,z) = \eta_1(n) + \eta_2(n,z) + \eta_3(n,z),$$

where

$$\begin{aligned}\eta_1(n) &:= -\,[\mathcal{S}^{-n}J\hat{V}, D^{-2}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)}]_{\mathbb{Z}(n)}, \\ \eta_2(n,z) &:= -\,[\mathcal{S}^{-n}J\hat{V}, D^{-2}T(Id-T)^{-1}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)}]_{\mathbb{Z}(n)}, \\ \eta_3(n,z) &:= -\,[\mathcal{S}^{-n}J\hat{V}, D^{-1}T(Id-T)^{-1}D^{-1}(Id-T)^{-1}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)}]_{\mathbb{Z}(n)}.\end{aligned}$$

The three terms η_1, η_2, and η_3 are estimated separately. The coefficient η_1 is independent of z and

$$\eta_1(n) = \frac{1}{\pi^4}\left(\frac{\hat{V}}{k^2} * \frac{\hat{V}}{k^2}\right)(2n) = \frac{1}{\pi^4}\sum_{k\neq\pm n}\frac{\hat{V}(n-k)}{(n-k)^2}\frac{\hat{V}(n+k)}{(n+k)^2}.$$

Hence, by Lemmas A.1 and A.2 (cf. formula (A.2) for $\alpha = 2$),

$$\left(\sum_{n\in\mathbb{Z}} w_2(2n)^2\eta_1(n)^2\right)^{1/2} \le \frac{1}{\pi^4}\left\|\frac{\hat{V}}{k^2} * \frac{\hat{V}}{k^2}\right\|_{w_2} \le \frac{6}{\pi^4}\left\|\frac{\hat{V}}{k^2}\right\|_{w_2}^2 \le \frac{6}{\pi^4}\|\hat{V}\|_w^2. \tag{2.33}$$

To estimate $\eta_2(n,z)$, introduce, for $|z| \le M, |n| \ge n_w$, and $k \in \mathbb{Z}\backslash\{-2n, 0\}$,

$$a(k,z) \equiv a_{(n)}(k,z) := -\mathcal{S}^{-n}T(Id-T)^{-1}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)}(k).$$

Then $\|T\|_{\mathcal{L}(\ell^2_{\mathcal{S}^n_w})} \le \frac{n_w}{3|n|}$ (Lemma 2.7) and $\|(Id-T)^{-1}\|_{\mathcal{L}(\ell^2_{\mathcal{S}^n_w})} \le 2$ (Corollary 2.8); hence

$$\|a_{(n)}\|_w \le \frac{n_w}{3|n|}2\|V\|_w,$$

and, as $w_2(n) = (1+|n|)^2w(2n)$ is submultiplicative,

$$\begin{aligned}(1+|n|)^2w(2n)|\eta_2(n,z)| &\le \frac{1}{\pi^4}\left|\frac{\hat{V}}{k^2} * \frac{a_{(n)}}{k^2}\right|(2n)\\ &\le \frac{1}{\pi^4}\sum_{k\neq\pm n}\frac{(1+|n-k|)^2}{|n-k|^2}w(n-k)|\hat{V}(n-k)|\frac{(1+|n+k|)^2}{|n+k|^2}|a_{(n)}(n+k)|\\ &\le \frac{4^2}{\pi^4}\|V\|_w\|a_{(n)}\|_w \le \frac{4^2}{\pi^4}\frac{2}{3}\frac{n_w}{|n|}\|V\|_w^2\end{aligned}$$

and

$$\begin{aligned}&\left(\sum_{|n|\ge n_w}(1+|n|)^4w(2n)^2\sup_{|z|\le M}|\eta_2(n,z)|^2\right)^{1/2}\\ &\le \frac{4^2}{\pi^4}\frac{2}{3}\|V\|_w^2\left(n_w^2\sum_{|n|\ge n_w}\frac{1}{n^2}\right)^{1/2} \le \frac{4^2}{\pi^4}\frac{2}{3}(2(1+n_w))^{1/2}\|V\|_w^2 \le \frac{4^2}{\pi^4}(1+n_w)^{1/2}\|V\|_w^2.\end{aligned} \tag{2.34}$$

Hence, we used that $N^2\sum_{|n|\ge N}\frac{1}{n^2} \le 2N^2(\frac{1}{N^2} + \int_N^\infty\frac{dx}{x^2}) = 2(1+N)$. To estimate $\eta_3(n,z)$, we proceed in the same way as for $\eta_2(n,z)$. Introduce, for $|z| \le M, |n| \ge n_w$, and $k \in \mathbb{Z}\backslash\{-2n, 0\}$,

$$a(k,z) \equiv a_{(n)}(k,z) = \mathcal{S}^{-n}T(Id-T)^{-1}D^{-1}(Id-T)^{-1}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)}(k).$$

Then, by Lemma 2.7 and Corollary 2.8,

$$\|a_{(n)}\|_w \le \frac{n_w}{3|n|} 2 \frac{1}{\pi^2|n|} 2\|V\|_w,$$

where we use that $|D_{kk}^{-1}| = \frac{1}{\pi^2|n^2-k^2|} \le \frac{1}{\pi^2|n|}$ $\forall k \in \mathbb{Z}(n)$. As w_2 is submultiplicative,

$$\begin{aligned}
&(1+|n|)^2 w(2n)|\eta_3(n,z)| \le \frac{1+|n|}{\pi^2} \left| \frac{\hat{V}}{k} * \frac{a(n)}{k} \right| (2n) \\
&\le \frac{1+|n|}{\pi^2} \sum_{k \neq \pm n} \frac{1+|n-k|}{|n-k|} w(n-k)|\hat{V}(n-k)| \frac{1+|n-k|}{|n+k|} w(n+k)|a_{(n)}(n+k)| \\
&\le \frac{4}{\pi^2}(1+|n|)\|V\|_w \|a_{(n)}\|_w \\
&\le \frac{4^2}{\pi^4} \frac{1}{3} \frac{|n|+1}{|n|} \frac{n_w}{|n|} \|V\|_w^2
\end{aligned}$$

and, with

$$\begin{aligned}
&\left(\sum_{|n| \ge n_w} \left(\frac{1+|n|}{n^2} \right)^2 \right)^{1/2} \le \sqrt{2} \left(\left(\frac{1+n_w}{n_w^2} \right)^2 + \int_{n_w}^{\infty} \left(\frac{1}{x^4} + \frac{2}{x^3} + \frac{1}{x^2} \right) dx \right)^{1/2} \\
&\quad \le 4 n_w^{-1/2},
\end{aligned}$$

one obtains

$$\begin{aligned}
&\left(\sum_{|n| \ge n_w} (1+|n|)^4 w(2n)^2 \sup_{|z| \le M} |\eta_3(n,z)|^2 \right)^{1/2} \\
&\quad \le \frac{4^3}{\pi^4} \frac{1}{3} \frac{4}{\sqrt{n_w}} n_w \|V\|_w^2 \le \frac{4^4 n_w^{1/2}}{3\pi^4} \|V\|_w^2.
\end{aligned} \tag{2.35}$$

Combining (2.34)–(2.35), we get

$$\begin{aligned}
&\left(\sum_{|n| \ge n_w} (1+|n|)^4 w(2n)^2 \sup_{|z| \le M} \left| \frac{d}{dz} \beta(n,z) \right|^2 \right)^{1/2} \\
&\le \frac{6 + 4^2 + 4^{4/3}}{\pi^4} (n_w+1)^{1/2} \|V\|_w^2 \le 2(n_w+1)^{1/2} \|V\|_w^2. \qquad \square
\end{aligned}$$

The previous lemmas lead us to the following main result of this section.

PROPOSITION 2.14. *The following statements hold:*

(i) $\left(\sum_{|n| \ge n_w} (1+|n|)^4 w(2n)^2 \sup_{|z| \le M} |\beta_2(n,z)|^2 \right)^{1/2} \le 2(1+n_w)^{3/2} \|V\|_w^2,$

(ii) $\left(\sum_{|n| \ge n_w} (1+|n|)^2 w(2n)^2 |\beta_1(n)|^2 \right)^{1/2} \le \|V\|_w^2,$

(iii) $\left(\sum_{|n| \ge n_w} (1+|n|)^2 w(2n)^2 \sup_{|z| \le M} |\beta(n,z)|^2 \right)^{1/2} \le 2(1+n_w)^{3/2} \|V\|_w^2.$

Proof. (i) By (2.23), for $|z| \le M$,

$$|\beta_2(n,z)| \le |\beta_2(n,0)| + M \sup_{|z| \le M} \left| \frac{d}{dz} \beta(n,z) \right|$$

with $M + \|V\|_w = n_w$

$$\begin{aligned}&\left(\sum_{|n|\ge n_w}(1+|n|)^4 w(2n)^2 \sup_{|z|\le M}|\beta_2(n,z)|^2\right)^{1/2}\\ &\le \|V\|_w^3 + 2M(1+n_w)^{1/2}\|V\|_w^2\\ &\le \|V\|_w^2(\|V\|_w + 2M)(1+n_w)^{1/2};\end{aligned}$$

(ii) Lemma 2.10;

(iii) It follows from conditions (i) and (ii). □

2.7. The ζ-equation. In this section, we analyze the ζ-equation, stated in (2.18),

$$\zeta^2 - \left(\hat{V}(2n) + \beta(n, z(\zeta))\right)\left(\hat{V}(-2n) + \beta(-n, z(\zeta))\right) = 0. \tag{2.36}$$

Given $M > 0$ and $V \in L_0^2$, define

$$r_n := \max_{\varepsilon=\pm1}|\hat{V}(\varepsilon 2n)| + \max_{\varepsilon=\pm1}|\beta_1(\varepsilon n)| + \max_{\substack{|z|\le M\\ \varepsilon=\pm1}}|\beta_2(\varepsilon n, z)|, \tag{2.37}$$

and let $n_* := M + \|V\|$. By Lemmas 2.10 and 2.11, applied for the weight $w = 1$,

$$r_n \le \|V\| + 2\|V\|^2 \quad \forall n \text{ with } |n| \ge n_*.$$

Proposition 2.15. *Assume that $M > 0$ satisfies*

$$3(1+\|V\|)^2 \le \frac{M}{4}. \tag{2.38}$$

Then, for $n \ge n_*, \zeta$-equation (2.36) *has exactly two (counted with multiplicity) solutions in the disc $\overline{\mathcal{D}_{r_n}}$.*

Notation. We label these two solutions by ζ_n^+, ζ_n^- in an arbitrary way, but then we keep them fixed.

Proof. Clearly, $\zeta^2 = 0$ has precisely two roots in any disc $\mathcal{D}_r$. For $|\zeta| = Kr_n$ with $1 < K < 2$ close to 1 and any $n \ge n_*$, by (2.37) and (2.23),

$$\sup_{|z|\le M}|\left(\hat{V}(2n) + \beta(n,z)\right)\left(\hat{V}(-2n) + \beta(-n,z)\right)| \le r_n^2 < |\zeta|^2 \tag{2.39}$$

and $|\zeta| < 2r_n \le \frac{M}{2}$. Taking into account (2.38), it follows from Proposition 2.6 that $z_n(\zeta) \in \mathcal{D}_M$ depends analytically on ζ for $|\zeta| < M/2$ and $n \ge n_*$. Therefore, the left side of (2.36), denoted by $g(\zeta)$, is an analytic function of ζ in $\mathcal{D}_{M/2}$ and, by (2.39),

$$g(\zeta) = \zeta^2 + g_1(\zeta)\ ; \quad |g_1(\zeta)| < |\zeta|^2 \quad \text{for } |\zeta| \le \frac{M}{2}.$$

Therefore, by Rouché's theorem, (2.36) has precisely two roots in $\mathcal{D}_{Kr_n}$. As the two roots are independent of K, and $1 < K < 2$ is arbitrarily close to 1, we conclude that $\zeta_n^\pm \in \overline{\mathcal{D}_{r_n}}$. □

Let, for $n \ge n_*$,

$$z_n^\pm := z(\zeta_n^\pm) = \zeta_n^\pm + \alpha(n, z(\zeta_n^\pm)), \tag{2.40}$$

where $\zeta_n^\pm$ are the two solutions of (2.36), given by Proposition 2.15, and define

$$\lambda_n^\pm := n^2\pi^2 + z_n^\pm. \tag{2.41}$$

Then $\lambda_n^\pm$ is a pair of periodic eigenvalues, $\lambda_n^\pm \in spec_{per}(-\frac{d^2}{dx^2}+V)$. In the next section we want to deduce estimates for the gap length sequence $(\lambda_n^+ - \lambda_n^-)_{n\geq 1}$.

2.8. Gap length estimates.

PROPOSITION 2.16. *Assume that $M > 0$ satisfies*

$$3(1+\|V\|_w)^2 \leq \frac{M}{4}.$$

Then, with $n_w := M + \|V\|_w$,

$$\left(\sum_{n\geq n_w} w(2n)^2|\lambda_n^+ - \lambda_n^-|^2\right)^{1/2} \leq 8(1+n_w)^{3/2}(1+\|V\|_w)^2.$$

Remark. With $N := 13(1+\|V\|_w)^2, K_1 = 500$ and $K_2 = 5$. Proposition 2.16 gives Theorem 1.1.

Proof. Notice that, for $n \geq n_w$, by (2.19),

$$|\lambda_n^+ - \lambda_n^-| = |z_n^+ - z_n^-| \leq |\zeta_n^+ - \zeta_n^-| + \sup_{|z|\leq M}\left|\frac{d}{dz}\alpha(n,z)\right| |z_n^+ - z_n^-|. \tag{2.42}$$

By Lemma 2.5, (as $|z| + \|V\| \leq M + \|V\|_w = n_w \leq n$),

$$\left|\frac{d}{dz}\alpha(n,z)\right| \leq \frac{\|V\|^2}{9n^2} \leq \frac{\|V\|^2}{9n_w^2} \leq \frac{1}{9}. \tag{2.43}$$

Substituting the estimate (2.43) into (2.42) yields

$$\frac{1}{2}|z_n^+ - z_n^-| \leq |\zeta_n^+ - \zeta_n^-|.$$

As $|\zeta_n^+ - \zeta_n^-| \leq |\zeta_n^+| + |\zeta_n^-| \leq 2r_n$, with r_n defined by (2.37), we then conclude that, for $n \geq n_w$,

$$|z_n^+ - z_n^-| \leq 4r_n.$$

In view of Proposition 2.14,

$$\left(\sum_{n\geq n_w} w(2n)^2 r_n^2\right)^{1/2} \leq 4\cdot 2(1+n_w)^{3/2}(1+\|V\|_w)^2. \qquad \square$$

2.9. Gap length asymptotics. In this section, we obtain the first two terms in the asymptotics of $\lambda_n^+ - \lambda_n^-$ for $n \to \infty$. Let

$$\begin{aligned} &\rho(\pm n) := \hat{V}(\pm 2n) + \beta_1(\pm n), \\ &\eta(z) \equiv \eta(n,z) := \beta_2(-n,z)\rho(n) + \beta_2(n,z)\rho(-n) + \beta_2(-n,z)\beta_2(n,z), \end{aligned} \tag{2.44}$$

where the decomposition $\beta(\pm n, z) = \beta_1(\pm n) + \beta_2(\pm n, z)$ has been defined in (2.24)–(2.25). Then (2.36) can be written as

$$\zeta^2 - \rho(n)\rho(-n) - \eta(z(\zeta)) = 0. \tag{2.45}$$

Lemma 2.17. *Assume that $M > 0$ satisfies $3(1 + \|V\|_w)^2 \leq \frac{M}{4}$. Then the following estimates hold:*

(i) $|\rho(n)| \leq \frac{M}{4} \quad \forall n$; $\|\varrho\|_w \leq \|V\|_w + \|V\|_w^2$;

(ii) $\left(\sum_{n \geq n_w} (1 + |n|)^2 w(2n)^2 \sup_{|z| \leq M} |\eta(n, z)|\right)^{1/2} \leq 4(1 + n_w)^{3/2}(1 + \|V\|_w)^2$.

Proof. (i) By Lemma 2.10, $|\beta_1(n)| \leq \|V\|^2$, and therefore,

$$|\rho(n)| \leq 2(1 + \|V\|)^2 \leq \frac{M}{4}.$$

Moreover, we have $\|\varrho\|_w < \|V\|_w + \|V\|_w^2$. (ii) By the definition of $\eta(n, z)$, and Proposition 2.14(i)

$$\begin{aligned}
&\sum_{n \geq n_w} (1 + |n|)^2 w(2n)^2 \sup_{|z| \leq M} |\eta(n, z)| \\
&\leq 2 \left(\sum_{|n| \geq n_w} (1 + |n|)^4 w(2n)^2 \sup_{|z| \leq M} |\beta_2(n, z)|^2 \right)^{1/2} 2(1 + \|V\|_w)^2 \\
&+ \sum_{n \geq n_w} (1 + |n|)^2 w(2n)^2 \sup_{|z| \leq M} |\beta_2(n, z)|^2 \\
&\leq 4(1 + \|V\|_w)^2 \cdot 2(1 + n_w)^{3/2}(1 + \|V\|_w)^2 + 4(1 + n_w)^3(1 + \|V\|_w)^4 \\
&\leq 12(1 + n_w)^3(1 + \|V\|_w)^4. \qquad \square
\end{aligned}$$

Lemma 2.18. *Assume that $M > 0$ satisfies $3(1 + \|V\|_w)^2 \leq \frac{M}{4}$. Then, for $n \geq n_w$, either of the two roots $\hat{\zeta} \in \{\zeta_n^\pm\}$ satisfies*

$$\min_{\pm} |\hat{\zeta} \pm (\rho(n)\rho(-n))^{1/2}| \leq 5 \sup_{|z| \leq M} |\eta(z)|^{1/2}.$$

Proof. Choose an arbitrary root R of $R^2 = \rho(n)\rho(-n)$ and let $s := \sup_{|z| \leq M} |\eta(z)|$. We distinguish the following two cases.

Case 1: $|R^2| \leq 4s$: we have, with $\hat{z} = z_n^\pm$ for $\hat{\zeta} = \zeta_n^\pm$,

$$\begin{aligned}
|(\hat{\zeta} \pm R)^2| &\leq 2|\hat{\zeta}^2| + 2|R^2| \\
&\leq 2|R^2| + 2|\eta(\hat{z})| + 2|R^2| \\
&\leq 4|R^2| + 2|\eta(\hat{z})| \leq 18s \leq (5s^{1/2})^2.
\end{aligned}$$

Case 2: $|R^2| > 4s$: in this case, $|R^2| > 0$ and (2.45) can be rewritten as

$$\zeta^2 = R^2 \left(1 + \frac{\eta(z(\zeta))}{R^2}\right), \tag{2.46}$$

where $z(\zeta)$ is a solution of the z-equation (2.17). Let $\xi := \frac{\zeta}{R}$. Then, (2.46) can be written as

$$\xi^2 = 1 + \frac{\eta(z(\zeta))}{R^2}. \tag{2.47}$$

By assumption, $|R^2| > 4s$ and, as $|z(\zeta)| \leq M$, $|\eta(z(\zeta))/R^2| \leq \frac{1}{4}$. Denoting by $(1+w)^{1/2}$ the branch of the square root determined by $1^{1/2} = 1$, we obtain the equations

$$\xi = F_{\pm}(\xi), \tag{2.48}$$

where $F_{\pm}(\xi) := \pm(1 + \frac{\eta(z)}{R^2})^{1/2}$, with $z \equiv z(R\xi)$. Let us first consider the equation $\xi = F_+(\xi)$. Let $\mathcal{D}_{\frac{1}{4}}(1) := \{\xi \in \mathbb{C} \quad | \quad |\xi - 1| < \frac{1}{4}\}$ and notice that, for $\xi \in \overline{\mathcal{D}_{\frac{1}{4}}(1)}, \zeta = R\xi$ satisfies $|\zeta| \leq \frac{M}{4}\frac{5}{4} < \frac{M}{2}$, where we used the estimate $|R| \leq \frac{M}{4}$ of Lemma 2.17(i). According to Proposition 2.6, $z = \zeta + \alpha(n, z)$ has a unique solution $z(\zeta) \in \mathcal{D}_M$. This shows that $F_+(\xi) = (1 + \frac{\eta(z(R\xi))}{R^2})^{1/2}$ is well defined for $\xi \in \overline{\mathcal{D}_{1/4}(1)}$.

As $|(1+x)^{1/2} - 1| \leq \frac{2}{3}|x|$ for $x \in \overline{\mathcal{D}_{1/4}(0)}$ and $|R^2| > 4s$, we conclude that F_+ maps $\overline{\mathcal{D}_{1/4}(1)}$ into itself. Furthermore, F_+ is continuous, and therefore, by Brower's fixed point theorem, $\xi = F_+(\xi)$ admits at least one fixed point $\xi^I \in \overline{\mathcal{D}_{1/4}(1)}$,

$$\xi^I = F_+(\xi^I) = \left(1 + \frac{\eta(z^I)}{R^2}\right)^{1/2},$$

where $z^I = z(R\xi^I)$ and ξ^I satisfies the estimate

$$|\xi^I - 1| \leq \left|\left(1 + \frac{\eta(z^I)}{R^2}\right)^{1/2} - 1\right| \leq \frac{2}{3}\left|\frac{\eta(z^I)}{R^2}\right| \leq \frac{2}{3} \cdot \frac{1}{2}\frac{s^{1/2}}{|R|},$$

where, for the last inequality, we used that $|R^2| > 4s$. Hence, $\zeta^I := R\xi^I$ satisfies

$$|\zeta^I - R| \leq \frac{1}{2}\sup_{|z| \leq M}|\eta(z)|^{1/2} = \frac{1}{2}s^{1/2}.$$

The same arguments can be used to show that there exists a solution $\xi^{II} \in \overline{\mathcal{D}_{1/4}(-1)}$ of the equation $\xi = F_-(\xi)$ so that $\zeta^{II} := R\xi^{II}$ satisfies

$$|\zeta^{II} + R| \leq \frac{1}{2}\sup_{|z| \leq M}|\eta(z)|^{1/2} = \frac{1}{2}s^{1/2}.$$

Therefore, with $2|R| \geq 4s^{1/2}$

$$\begin{aligned} |\zeta^I - \zeta^{II}| &= |2R - (R - \zeta^I) - (\zeta^{II} + R)| \\ &\geq 2|R| - \frac{1}{2}S^{1/2} - \frac{1}{2}S^{1}2 \geq 3S^{1/2} > 0; \end{aligned} \tag{2.49}$$

hence, $\zeta^I \neq \zeta^{II}$. Moreover, ζ^I and ζ^{II} are solutions of (2.45) and thus satisfy, in view of (2.36),

$$|\zeta^I|, |\zeta^{II}| \leq r_n := \max_{\pm}|\hat{V}(\pm 2n)| + \max_{\pm}|\beta_1(\pm n)| + \max_{|z| \leq M}|\beta_2(\pm n, z)|.$$

Therefore, by Proposition 2.15, $\{\zeta^I, \zeta^{II}\} = \{\zeta_n^+, \zeta_n^-\}$. □

For later use, we state the following application of Lemma 2.18.

COROLLARY 2.19. *Let $V \in H_0^w$ be a 1-periodic potential. Then for M with $3(1+\|V\|_w)^2 \leq \frac{M}{4}$,*

$$\left(\sum_{h\geq n_w} w(2n)^2|\zeta_n^\pm|^2\right)^{1/2} \leq 9(1+n_w)^{3/2}(1+\|V\|_w)^4.$$

Proof. By (2.45)

$$\begin{aligned}&\left(\sum_{n\geq n_w} w(2n)^2|\zeta_n^\pm|^2\right)^{1/2}\\ &\leq \left(\sum_{n\geq n_w} w(2n)^2|\rho(n)\rho(-n)|\right)^{1/2} + \left(\sum_{n\geq n_w} w(2n)^2 \sup_{|z|\leq M}|\eta(n,z)|\right)^{1/2}.\end{aligned}$$

By Lemma 2.17,

$$\left(\sum_{n\geq n_w} w(2n)^2 \sup_{|z|\leq M}|\eta(n,z)|\right)^{1/2} \leq 4(1+n_w)^{3/2}(1+\|V\|_w)^2$$

and, with $\rho(n) = \hat{V}(2n) + \beta_1(n)$,

$$\begin{aligned}&\left(\sum_{n\geq n_w} w(2n)^2|\rho(n)\rho(-n)|\right)^{1/2}\\ &\leq \left(\sum_{n\geq n_w} w(2n)^2|\rho(n)|^2\right)^{1/2}\left(\sum_{n\geq n_w} w(2n)^2|\rho(-n)|^2\right)^{1/2}\\ &\leq (\|V\|_w + \|V\|_w^2)^2 \leq (1+\|V\|_w)^4,\end{aligned}$$

where we have used Lemma 2.17. □

Recall that $\lambda_n^\pm = n^2\pi^2 + z_n^\pm$ denote periodic eigenvalues of the operator $-\frac{d^2}{dx^2}+V$ and $\rho(\pm n)$ have been defined in (2.44).

THEOREM 2.20. *Let $V \in H_0^w$ be 1-periodic. Then, for any $M > 0$ with $3(1+\|V\|_w)^2 \leq \frac{M}{4}$ and $n_w = M + \|V\|_w$,*

$$\begin{aligned}&\left(\sum_{n\geq n_w}(1+|n|)^2 w(2n)^2\left(\min_{\pm}\left((\lambda_n^+-\lambda_n^-)\pm 2(\rho(n)\rho(-n))^{1/2}\right)\right)^2\right)^{1/2}\\ &\leq 50(1+n_w)^{3/2}(1+\|V\|_w)^4. \qquad (2.50)\end{aligned}$$

Remark. With $N := 13(1+\|V\|_w)^2, K_3 := 10^6$, and $K_4 := 14$, Theorem 2.20 gives Theorem 1.2.

Proof. For $n \geq n_w, \lambda_n^+ - \lambda_n^- = z_n^+ - z_n^-$. Furthermore, $z_n^\pm = \zeta_n^\pm + \alpha(n, z_n^\pm)$ and, by

Lemma 2.18 $\min_\pm|(\zeta_n^+-\zeta_n^-)\pm 2(\rho(n)\rho(-n))^{1/2}| \le 10\sup_{|z|\le M}|\eta(n,z)|^{1/2}$. Therefore,

$$\begin{aligned}&\min_\pm|(\lambda_n^+-\lambda_n^-)\pm 2(\rho(n)\rho(-n))^{1/2}|\\&\le \min_\pm|(\zeta_n^+-\zeta_n^-)\pm 2(\rho(n)\rho(-n))^{1/2}| + \left(\sup_{|z|\le M}\left|\frac{d}{dz}\alpha(n,z)\right|\right)|z_n^+-z_n^-|\\&\le 10\sup_{|z|\le M}|\eta(n,z)|^{1/2}+\frac{\|V\|^2}{n^2}|z_n^+-z_n^-|,\end{aligned}$$

where for the last inequality we used Lemma 2.5(ii). By Lemma 2.17(ii) (estimate for $\sup_{|z|\le M}|\eta(n,z)|^{1/2}$) and by Proposition 2.16 (estimate for $|z_n^+-z_n^-|=|\lambda_n^+-\lambda_n^-|$),

$$\begin{aligned}&\left(\sum_{n\ge n_w}(1+|n|)^2w(2n)^2\min_\pm|(\lambda_n^+-\lambda_n^-)\pm 2(\rho(n)\rho(-n))^{1/2}|^2\right)^{1/2}\\&\le 10\left(4(1+n_w)^{3/2}(1+\|V\|_w)^2\right)\\&+\|V\|^2\left(\sum_{n\ge n_w}\frac{(1+n)^2}{n^4}w(2n)^2|z_n^+-z_n^-|^2\right)^{1/2}\\&\le 40(1+n_w)^{3/2}(1+\|V\|_w)^2+\|V\|_w^2 8(1+n_w)^{3/2}(1+\|V\|_w)^2\\&\le 50(1+n_w)^{3/2}(1+\|V\|_w)^4. \qquad \square\end{aligned}$$

3. Eigenfunctions and Riesz's spaces.

3.1. Eigenfunctions. In this section we review the estimates of the Fourier coefficients of an L_2-normalized eigenfunction f corresponding to a periodic eigenvalue $\lambda = n^2\pi^2+z$ of $L=-\frac{d^2}{dx^2}+V$. f is a 2-periodic function in $H^2_{loc}(\mathbb{R};\mathbb{C})$ satisfying

$$(L-\lambda)f=0;\ \|f\|=1.$$

Recall that $x^f := \hat{f}(-n), y^f := \hat{f}(n)$, and $F := (\hat{f}(k))_{k\in\mathbb{Z}(n)}$. By Proposition 2.2, $F = x^fF_+ + y^fF_-$ and, by (2.22), $F_\pm := (z-A_n)^{-1}(\mathcal{S}^{\pm n}\hat{V})_{\mathbb{Z}(n)}$. For $n\ge n_w$, F_+ and F_- satisfy the estimates (cf. Corollary 2.9)

$$\|F_+\|_{\ell^2_{\mathcal{S}^n w}} \le \frac{2}{\pi^2 n}\|V\|_w, \tag{3.1}$$

$$\|F_-\|_{\ell^2_{\mathcal{S}^{-n} w}} \le \frac{2}{\pi^2 n}\|V\|_w. \tag{3.2}$$

By the normalization of f

$$1=\int_0^2 f(x)\overline{f(x)}dx = 2\left(|x^f|^2+|y^f|^2+\sum_{k\ne\pm n}|\hat{f}(k)|^2\right). \tag{3.3}$$

In particular, one has

$$|x^f|^2+|y^f|^2\le\frac{1}{2}. \tag{3.4}$$

Hence, by Cauchy's inequality $|F(k)| \leq (|x^f|^2 + |y^f|^2)^{1/2}(|F_+(k)|^2 + |F_-(k)|^2)^{1/2}$, we obtain in view of (3.1)–(3.2)

$$\|F\|_{\ell^2(\mathbb{Z}(n))} \leq \left(\frac{1}{2} \cdot 2 \cdot \frac{2}{\pi^2 n}\|V\|\right)^{1/2} \leq \frac{1}{4n}\|V\|_w.$$

Thus, for n with $n \geq n_w$,

$$\|F\|_{\ell^2(\mathbb{Z}(n))} \leq \frac{1}{4}.$$

Together with (3.3)–(3.4), this yields

$$\frac{1}{4} \leq |x^f|^2 + |y^f|^2 \leq \frac{1}{2}. \tag{3.5}$$

We summarize our estimates as follows.

Lemma 3.1. *Let $V \in H_0^w$ be 1-periodic. Then for any $M > 0$ and $n \geq n_w$ with*

$$3(1 + \|V\|_w)^2 \leq \frac{M}{4}; \; n \geq n_w := M + \|V\|_w,$$

an eigenfunction f with $\|f\| = 1$, corresponding to an eigenvalue λ with $|\lambda - n^2\pi^2| \leq M$, has the following properties:

(i) $f(x) = x^f e^{-in\pi x} + y^f e^{in\pi x} + x^f F_+ + y^f F_-$;
(ii) $\frac{1}{4} \leq |x^f|^2 + |y^f|^2 \leq \frac{1}{2}$;
(iii) $\|F_+\|_{\ell^2_{\mathcal{S}^{n_w}}} \leq \frac{\|V\|_w}{4n}$; $\|F_-\|_{\ell^2_{\mathcal{S}^{-n_w}}} \leq \frac{\|V\|_w}{4n}$.

3.2. Riesz's spaces. Given a 1-periodic potential $V \in H_0^w$, let $M > 0$ satisfy $3(1 + \|V\|_w)^2 \leq \frac{M}{4}$. For $n \geq n_w := M + \|V\|_w$, there are precisely two (counted with multiplicity) eigenvalues, λ_n^+ and λ_n^-, near $n^2\pi^2$ of $L = -\frac{d^2}{dx^2} + V$. Recall that $specL = specL_{Per^+} \cup specL_{Per^-}$. Denote by $P_n^{\pm}$ the Riesz projectors corresponding to the boundary conditions $Per^{\pm}$ (cf. (1.16)) and let

$$E_{2n} := P_n^+(L^2[0,1]); \quad E_{2n-1} := P_n^-(L^2[0,1]), \quad (\forall n \geq 1).$$

If $\lambda_n^+ \neq \lambda_n^-$ or $\lambda_n^+ = \lambda_n^-$ is of geometric multiplicity 2, there exist two linearly independent eigenfunctions, corresponding to the eigenvalues λ_n^+ and λ_n^-, and E_n is given by the linear span of these two eigenfunctions. In the case where $\lambda_n^+ = \lambda_n^-$ is of geometric multiplicity 1, E_n denotes the root space of λ_n^+. Notice that this case might happen if the potential V is complex-valued. As an example we mention $V = \varepsilon e^{2\pi ix}$ ($\varepsilon \neq 0$ arbitrary). The periodic eigenvalues of $-\frac{d^2}{dx^2} + \varepsilon e^{2\pi ix}$ (considered on the interval $[0,2]$) are given by $n^2\pi^2$ ($n \geq 0$), where for every $n \geq 1, n^2\pi^2$ is a double eigenvalue of geometric multiplicity 1 (cf. [3], [5] for details).

Let us describe E_n in the case where $\lambda_n^+ = \lambda_n^-$ is of geometric multiplicity 1 in more detail. Denote by f an L_2-normalized eigenfunction corresponding to $\lambda_n^+ = \lambda_n^-$, $Lf = \lambda_n^+ f$. Choose an L_2-normalized element φ in E_n, orthogonal to f. Then $E_n = span(f, \varphi)$ and φ satisfies

$$(L - \lambda_n^+)\varphi = \xi_n f$$

with $\xi_n \neq 0$. Denote by $c(x,\lambda)$ and $s(x,\lambda)$ the fundamental solution of $-y'' + Vy = \lambda y$ with

$$c(0,\lambda) = 1, \quad c'(0,\lambda) = 0; \quad s(0,\lambda) = 0, \quad s'(0,\lambda) = 1. \tag{3.6}$$

Lemma 3.2. *Assume $\lambda_{2n} = \lambda_{2n-1}$ and $\int_0^1 s(x, \lambda_{2n})^2 dx \neq 0$. Then λ_{2n} is of geometric multiplicity 2 iff λ_{2n} is a Dirichlet eigenvalue of the operator L on $[0, 1]$.*

Proof. Assume that $\lambda \equiv \lambda_{2n}$ is of geometric multiplicity 2. Then the fundamental solutions $c(x, \lambda)$ and $s(x, \lambda)$ are eigenfunctions, both either periodic or antiperiodic, and $s(0, \lambda) = 0$. It follows that $s(1, \lambda) = 0$, and therefore, λ is a Dirichlet eigenvalue. Conversely, assume that $\lambda \in spec_{Dir}(L)$ is a double periodic eigenvalue. Then $\Delta(\lambda) := c(1, \lambda) + s'(1, \lambda) = \pm 2$ and $\dot{\Delta}(\lambda) := \frac{d}{d\lambda}\Delta(\lambda) = 0$, as well as

$$s(1, \lambda) = 0. \tag{3.7}$$

By the Wronskian identity,

$$1 = c(1, \lambda)s'(1, \lambda) - c'(1, \lambda)s(1, \lambda) = c(1, \lambda)s'(1, \lambda)$$

and, combined with $\Delta(\lambda) = \pm 2$, one obtains

$$c(1, \lambda) = s'(1, \lambda) = \pm 1. \tag{3.8}$$

Take the derivative of the Wronskian identity with respect to λ and use $\dot{\Delta}(\lambda) = 0$ to conclude that

$$\begin{aligned} 0 = {} & \dot{c}(1, \lambda)c'(1, \lambda) + c(1, \lambda)\dot{s}'(1, \lambda) \\ & -\dot{c}'(1, \lambda)s(1, \lambda) - c'(1, \lambda)\dot{s}(1, \lambda) \\ = {} & \pm\left(\dot{c}(1, \lambda) + \dot{s}'(1, \lambda)\right) - c'(1, \lambda)\dot{s}(1, \lambda) \\ = {} & 0 - c'(1, \lambda)\dot{s}(1, \lambda). \end{aligned}$$

As $\lambda \in spec_{Dir}(L)$, and $\int_0^1 s(x, \lambda)^2 dx \neq 0, \dot{s}(1, \lambda) \neq 0$, and therefore,

$$c'(1, \lambda) = 0, \tag{3.9}$$

i.e., λ is a Neumann eigenvalue of the operator L on $[0, 1]$. By (3.6)–(3.9), $c(x, \lambda)$ and $s(x, \lambda)$ are both periodic eigenfunctions of L on $[0, 2]$; hence, λ has geometric multiplicity 2. □

3.3. Orthonormal basis of E_n. In this section we obtain properties for an orthonormal basis f, φ of the two-dimensional subspace E_n introduced above ($n \geq n_w$, where $n_w := M + \|V\|_w$). Here f is an eigenfunction of $L = -\frac{d^2}{dx^2} + V$, with $\|f\|_{L^2} = 1$, corresponding to the eigenvalue $\lambda^+ \equiv \lambda_n^+$

$$Lf = \lambda^+ f \tag{3.10}$$

and φ is an element in E_n with

$$\langle \phi, f \rangle = 0; \quad \|\varphi\|_{L^2} = 1. \tag{3.11}$$

Notice that φ is determined up to a scalar $\kappa \in \{z \in \mathbb{C} \mid |z| = 1\}$. Here $\langle p, q \rangle$ denotes the L_2-inner product

$$\langle p, q \rangle = \int_0^2 p(x)\overline{q(x)} dx.$$

In the case when $\lambda^+ \equiv \lambda_n^+$ is a double eigenvalue, φ satisfies an equation of the form

$$L\varphi = \lambda^+ \varphi + \xi f \quad (\lambda^+ = \text{ double eigenvalue}), \tag{3.12}$$

where $\xi \equiv \xi_n = 0$ if λ_n^+ has geometric multiplicity 2 and $\xi \neq 0$ if λ^+ has geometric multiplicity 1. In the case where $\lambda^- \equiv \lambda_n^- \neq \lambda_n^+$, choose a normalized eigenfunction $f^- \equiv f_n^-$ of λ^- such that the following holds:

$$0 \leq a := \langle f^-, f \rangle \leq 1; \quad \|f^-\|_{L^2} = 1. \tag{3.13}$$

Write f^- as a linear combination of f and φ,

$$f^- = af + b\varphi, \tag{3.14}$$

where now the scalar κ for φ (cf. (3.11)) is chosen in such a way that $0 \leq b$. Then $a^2 + b^2 = 1$, and $b \neq 0$, or

$$a = \cos\theta; \quad b = \sin\theta; \quad 0 < \theta \leq \pi/2. \tag{3.15}$$

The function $\varphi = \frac{1}{b} f^- - \frac{a}{b} f$ satisfies

$$\begin{aligned} L\varphi &= \lambda^- \frac{1}{b} f^- - \lambda^+ \frac{a}{b} f \\ &= \lambda^+ \left(\frac{1}{b} f^- - \frac{a}{b} f \right) + (\lambda^- - \lambda^+) \frac{1}{b} f^- \\ &= \lambda^+ \varphi + (\lambda^+ - \lambda^-) \frac{1}{b} (f - f^-) - (\lambda^+ - \lambda^-) \frac{1}{b} f. \end{aligned}$$

Thus, in the case $\lambda^+ \neq \lambda^-$, with $\lambda \equiv \lambda^+$,

$$L\varphi = \lambda\varphi + \xi f + \gamma h, \tag{3.16}$$

where $\gamma \equiv \gamma_n = \lambda_n^+ - \lambda_n^-, h = \frac{1}{b}(f - f^-)$, and $\xi \equiv \xi_n$ is defined by

$$\xi_n := -(\lambda_n^+ - \lambda_n^-) \frac{1}{b} \qquad (\text{case } \lambda_n^+ \neq \lambda_n^-). \tag{3.17}$$

Notice that (3.12) has the same form as (3.16) if we set h equal to 0. It turns out that we will no longer have to treat the following three cases separately.

Case 1. $\lambda^+ = \lambda^-$ *and* $\xi = 0$;

Case 2. $\lambda^+ = \lambda^-$ *and* $\xi \neq 0$;

Case 3. $\lambda^+ \neq \lambda^-$.

The next result shows that the term $\gamma h = (\lambda^+ - \lambda^-) \frac{1}{b}(f - f^-)$ in (3.16) is well under control.

LEMMA 3.3. *If $\lambda^+ \neq \lambda^-$, then $b \neq 0$ and $\|h\| \leq \sqrt{2}$.*

Proof. By (3.15), $b = sin\theta \neq 0$ for $\lambda^+ \neq \lambda^-$. By (3.14) and (3.15), $h = \frac{1}{b}(f - f^-) = \frac{1-\cos\theta}{\sin\theta} f - \varphi$, and therefore, as f and φ are orthogonal,

$$\|h\|^2 = \left| \frac{1 - \cos\theta}{\sin\theta} \right|^2 + 1 \leq 2$$

as $\frac{1-\cos\theta}{\sin\theta} = \frac{2\sin^2\frac{\theta}{2}}{2\cos\frac{\theta}{2}\sin\frac{\theta}{2}} = \tan\frac{\theta}{2} \leq 1$ for $0 < \theta \leq \frac{\pi}{2}$. □

In the remaining part of this section our aim is to obtain estimates for ξ_n (cf. (3.12) and (3.17)). To this end, we write (3.16) in Fourier space. Introduce

$$\varphi = x^\varphi e^{-in\pi x} + y^\varphi e^{in\pi x} + \sum_{k \neq \pm n} \Phi(k) e^{ik\pi x}; \quad \Phi = (\Phi(k))_{k \in \mathbb{Z}(n)},$$

and, similarly

$$h = x^h e^{-in\pi x} + y^h e^{in\pi x} + \sum_{k \neq \pm n} H(k) e^{ik\pi x}, \quad H := (H(k))_{k \in \mathbb{Z}(n)}. \tag{3.18}$$

In view of (2.2)–(2.4), (3.16) leads to the following inhomogeneous system:

$$-zx^\varphi + \hat{V}(-2n)y^\varphi + [\mathcal{S}^n J\hat{V}, \Phi]_{\mathbb{Z}(n)} = \xi_n x^f + \gamma_n x^h, \tag{3.19}$$

$$\hat{V}(2n)x^\varphi - zy^\varphi + [\mathcal{S}^{-n} J\hat{V}, \Phi]_{\mathbb{Z}(n)} = \xi_n y^f + \gamma_n y^h, \tag{3.20}$$

$$(\mathcal{S}^n \hat{V})_{\mathbb{Z}(n)} x^\varphi + (\mathcal{S}^{-n}\hat{V})_{\mathbb{Z}(n)} y^\varphi + (A_n - z)\Phi = \xi_n F + \gamma_n H, \tag{3.21}$$

where, as usual, $z \equiv z_n^+ = \lambda_n^+ - n^2\pi^2$.

We use this system to obtain an estimate for ξ_n. The sequence Φ is obtained from (3.21),

$$\begin{aligned} \Phi = {} & (z - A_n)^{-1}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)} x^\varphi + (z - A_n)^{-1}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)} y^\varphi \\ & - (z - A_n)^{-1}\xi_n F - (z - A_n)^{-1}\gamma_n H \end{aligned} \tag{3.22}$$

and, by (2.11), F is given by

$$F = (z - A_n)^{-1}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)} x^f + (z - A_n)^{-1}(\mathcal{S}^{-n}\hat{V})_{\mathbb{Z}(n)} y^f. \tag{3.23}$$

Hence,

$$\begin{aligned} \Phi = {} & (z - A_n)^{-1}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)} x^\varphi + (z - A_n)^{-1}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)} y^\varphi \\ & - (z - A_n)^{-2}(\mathcal{S}^n\hat{V})_{\mathbb{Z}(n)} \xi_n x^f - (z - A_n)^{-2}(\mathcal{S}^{-n}\hat{V})_{\mathbb{Z}(n)} \xi_n y^f \\ & - (z - A_n)^{-1}\gamma_n H. \end{aligned}$$

In this form, substitute Φ into (3.19)–(3.20) to obtain, with $\alpha(n,z)$ and $\beta(n,z)$ defined by (2.13) and (2.14),

$$\begin{aligned} & \begin{pmatrix} -z + \alpha(n,z) & \hat{V}(-2n) + \beta(-n,z) \\ \hat{V}(2n) + \beta(n,z) & -z + \alpha(n,z) \end{pmatrix} \begin{pmatrix} x^\varphi \\ y^\varphi \end{pmatrix} \\ & = \xi_n \begin{pmatrix} x^f - \frac{d}{dz}\alpha(n,z)x^f - \frac{d}{dz}\beta(-n,z)y^f \\ y^f - \frac{d}{dz}\beta(n,z)x^f - \frac{d}{dz}\alpha(n,z)y^f \end{pmatrix} \\ & \quad + \gamma_n \begin{pmatrix} x^h + [\mathcal{S}^n J\hat{V}, (z - A_n)^{-1}H] \\ y^h + [\mathcal{S}^{-n} J\hat{V}, (z - A_n)^{-1}H] \end{pmatrix}, \end{aligned}$$

where we used $\alpha(n,z) = \alpha(-n,z)$, Lemma 2.4, and (cf. Lemma 2.5)

$$\frac{d}{dz}\alpha(n,z) = -[\mathcal{S}^{-n} J\hat{V}, (z - A_n)^{-2}(\mathcal{S}^{-n}\hat{V})_{\mathbb{Z}(n)}]_{\mathbb{Z}(n)}$$

and

$$\frac{d}{dz}\beta(n,z) = -[\mathcal{S}^{-n} J\hat{V}, (z - A_n)^{-2}(\mathcal{S}^{n}\hat{V})_{\mathbb{Z}(n)}]_{\mathbb{Z}(n)}.$$

Therefore,

$$\xi_n \left(Id_2 - \tfrac{d}{dz} \begin{pmatrix} \alpha(n,z) & \beta(-n,z) \\ \beta(n,z) & \alpha(n,z) \end{pmatrix} \right) \begin{pmatrix} x^f \\ y^f \end{pmatrix} \tag{3.24}$$

$$= \begin{pmatrix} -\zeta_n^+ & \hat{V}(-2n) + \beta(-n,z) \\ \hat{V}(2n) + \beta(n,z) & -\zeta_n^+ \end{pmatrix} \begin{pmatrix} x^\varphi \\ y^\varphi \end{pmatrix}$$

$$-\gamma_n \begin{pmatrix} x^h + [\mathcal{S}^n J\hat{V}, (z - A_n)^{-1} H] \\ y^h + [\mathcal{S}^{-n} J\hat{V}, (z - A_n)^{-1} H] \end{pmatrix},$$

where $\zeta_n^+ = z_n^+ - \alpha(n, z_n^+), Id_2$ denotes the 2×2 identity matrix, and to simplify notation, $[\cdot,\cdot] = [\cdot,\cdot]_{\mathbb{Z}(n)}$. Let $V \in H_0^w$ be 1-periodic and choose $M > 0$ with $3(1 + \|V\|)^2 \leq \frac{M}{4}$. By Lemma 2.5, for any $n \geq n_w := M + \|V\|_w$ and $z = z_n^+$

$$\left| \frac{d}{dz} \alpha(n,z) \right| \leq \frac{1}{9} \frac{\|V\|^2}{n^2} \leq \frac{1}{9}, \tag{3.25}$$

where we use that, by Lemma 2.5 and Propositions 2.6 and 2.15, $|z_n^+| \leq M$. Further, by Lemma 2.13, applied to $w = 1$,

$$(1 + |n|)^2 \left| \frac{d}{dz} \beta(n,z) \right| \leq 2(1 + n_w)^{1/2} \|V\|^2, \quad (|n| \geq n_w, \quad |z| \leq M).$$

Use that for $|n| \geq n_w$, $\frac{\|V\|^2}{1+|n|} \leq \frac{M/12}{n_w} \leq \frac{1}{12}$. Thus, for $|n| \geq n_w$ and $z \equiv z_n^+$,

$$\left| \frac{d}{dz} \beta(n,z) \right| \leq \frac{1}{6}. \tag{3.26}$$

Combining (3.25) and (3.26), the left-hand side of (3.24) can be estimated from below, for $n \geq n_w$,

$$\begin{aligned} &\left\| \xi_n \left(Id_2 - \frac{d}{dz} \begin{pmatrix} \alpha(n,z) & \beta(-n,z) \\ \beta(n,z) & \alpha(n,z) \end{pmatrix} \right) \begin{pmatrix} x^f \\ y^f \end{pmatrix} \right\|^2 \\ &\qquad \geq |\xi_n|^2 \left(|x^f|^2 \left(1 - \frac{1}{10} \right) + |y^f|^2 \left(1 - \frac{1}{10} \right) \right) \geq \left(\frac{1}{3} |\xi_n| \right)^2, \end{aligned} \tag{3.27}$$

where we used that $\frac{1}{4} \leq |x^f|^2 + |y^f|^2$ (cf. Lemma 3.1). Further, we need an estimate for $[\mathcal{S}^{\pm} J\hat{V}, (z - A_n)^{-1} H]$ with H as in (3.18).

LEMMA 3.4. *For* $n \geq n_w$,

(i) $|x^h| \leq \sqrt{2}$; $|y^h| \leq \sqrt{2}$;

(ii) $|[\mathcal{S}^{\pm n} J\hat{V}, (z - A_n)^{-1} H]| \leq \frac{\|V\|}{\pi n}$.

Proof. The proof of (i) follows from Lemma 3.3. To prove (ii), notice that for $n \geq n_w$ and $z = z_n^+$,

$$\begin{aligned} &|[\mathcal{S}^{\pm n} J\hat{V}, (z - A_n)^{-1} H]| \\ &\qquad \leq \|V\| \|(z - A_n)^{-1}\| \|H\| \leq \frac{2}{\pi^2} \frac{1}{n} \|V\| \sqrt{2} \\ &\qquad \leq \frac{2}{\pi^2} \frac{1}{n} \|V\| \sqrt{2}, \end{aligned}$$

where we used that $\sqrt{2}\|H\| \leq \|h\| \leq \sqrt{2}$ (cf. (3.18), Lemma 3.3), and $\|(z-A_n)^{-1}\| \leq \frac{2}{\pi^2}\frac{1}{n}$ by Lemma 2.1. □

Lemma 3.4 enables us to obtain an estimate *from above* of the right-hand side of (3.24). Lemma 3.4 and estimate (3.27), together with $|x^\varphi|^2 + |y^\varphi|^2 \leq 1$, are used to deduce from (3.24) that, for $n \geq n_w = M + \|V\|_w$,

$$\begin{aligned}\frac{1}{3}|\xi_n| &\leq \Big(|\zeta_n^+| + |\hat{V}(2n)| + |\beta(n,z)|\Big) \\ &\quad + \Big(|\zeta_n^+| + |\hat{V}(-2n)| + |\beta(-n,z)|\Big) + |\gamma_n|\left(\sqrt{2} + \frac{2\|V\|}{\pi n}\right) \\ &\leq 2|\zeta_n^+| + |\hat{V}(2n)| + |\hat{V}(-2n)| + |\beta(n,z)| + |\beta(-n,z)| + 3|\gamma_n|.\end{aligned} \tag{3.28}$$

Estimate (3.28), combined with earlier estimates for $\beta(\pm n, z)$ and γ_n, leads to the following inequality:

$$|\xi_n| \leq \frac{C}{w(2n)} \qquad \forall n \geq n_w, \tag{3.29}$$

where $C > 0$ depends only on $\|V\|_w$. In fact, the following stronger statement holds.

THEOREM 3.5. *Let $V \in H_0^w$ be a 1-periodic potential and let $M > 0$ satisfy $3(1+\|V\|_w)^2 \leq \frac{M}{4}$. Then the sequence $(\xi_n)_{n\geq n_w}$ (cf. (3.12), (3.17)) satisfies*

$$\left(\sum_{n\geq n_w} w(2n)^2|\xi_n|^2\right)^{1/2} \leq 120(1+n_w)^2(1+\|V\|_w)^2$$

with $n_w := M + \|V\|_w$.

Proof. The terms on the right side of (3.28) are estimated separately. Recall that, by Corollary 2.19,

$$\left(\sum_{n\geq n_w} w(2n)^2|\zeta_n^+|^2\right)^{1/2} \leq 5(1+n_w)^2(1+\|V\|_w)^2, \tag{3.30}$$

by Proposition 2.14,

$$\left(\sum_{|n|\geq n_w} w(2n)^2 \sup_{|z|\leq M}|\beta(n,z)|^2\right)^{1/2} \leq 2(1+n_w)^2(1+\|V\|_w)^2, \tag{3.31}$$

and by Proposition 2.16, with $|\gamma_n| = |\lambda_n^+ - \lambda_n^-|$,

$$\left(\sum_{n\geq n_w} w(2n)^2|\gamma_n|^2\right)^{1/2} \leq 8(1+n_w)^2(1+\|V\|_w)^2. \tag{3.32}$$

Combining (3.30)–(3.32) with (3.28) leads to the following estimate:

$$\frac{1}{3}\left(\sum_{n\geq n_w} w(2n)^2|\xi_n|^2\right)^{1/2} \leq 40(1+n_w)^2(1+\|V\|_w)^2. \qquad \square$$

3.4. Restriction of L on E_n. Here we summarize the results of sections 3.1–3.3 as a statement on the structure of the restriction of L on the Riesz spaces E_n.

PROPOSITION 3.6. *Let $V \in H_0^w$ be a 1-periodic potential. Then, for n sufficiently large, the Riesz space E_n has an orthonormal basis $f \equiv f_n$, $\varphi \equiv \varphi_n$ such that*

$$L_{Per^\varepsilon} f = \lambda_{2n} f; \quad L_{Per^\varepsilon} \varphi = \lambda_{2n} \varphi + \xi_n f + \gamma_n h,$$

where $\varepsilon \in \{+,-\}$ is $+$ for n even and $-$ for n odd and $h \equiv h_n \in E_n$. Moreover, the following inequalities hold:

$$\|h\| \leq 2; \quad |\xi_n| \leq \frac{C}{w(2n)}; \quad |\gamma_n| \leq \frac{C}{w(2n)}$$

with $C > 0$ independent of n. (For stronger estimates, cf. (3.30) *and* (3.32).*)*

4. Dirichlet spectrum.

4.1. Candidates for Dirichlet eigenfunctions. In section 3.2 we have introduced, for n sufficiently large, the two-dimensional subspaces E_n,

$$E_{2n} = \text{Range}(P_n^+); \quad E_{2n-1} = \text{Range}(P_n^-). \tag{4.1}$$

We have chosen an orthonormal basis (f, φ) of E_n with f being a normalized eigenfunction for the eigenvalue $\lambda \equiv \lambda_n^+$,

$$Lf = \lambda f,$$

and we showed that φ satisfies an equation of the form

$$L\varphi = \lambda\varphi + \xi f + \gamma h, \tag{4.2}$$

where $\gamma \equiv \gamma_n = \lambda_{2n} - \lambda_{2n-1}, h$ satisfies $\|h\| \leq 2$ (cf. Lemma 3.3), and estimates for ξ have been established in Theorem 3.5. The following lemma gives an element G in E_n, satisfying Dirichlet boundary conditions.

LEMMA 4.1. *Assume that $3(1 + \|V\|_w)^2 \leq \frac{M}{4}$ and $n \geq n_w = M + \|V\|_w$. Then there exists an element G in E_n of the form*

$$G = \alpha f + \beta \varphi; \quad 0 \leq \alpha \leq 1; \quad |\alpha|^2 + |\beta|^2 = 1 \tag{4.3}$$

so that

$$G(0) = 0; \quad G(1) = 0.$$

Proof. First consider the case where $f(0) = 0$. Then, as f is either periodic or antiperiodic,

$$f(1) = \pm f(0) = 0. \tag{4.4}$$

Thus $G := f$ has the required properties. If $f(0) \neq 0$, notice that $\tilde{G}(x) := -f(0)\varphi(x) + \varphi(0)f(x)$ is a nonzero element in E_n, satisfying Dirichlet boundary conditions. Then $G := \kappa \frac{\tilde{G}}{\|\tilde{G}\|} = \alpha f + \beta \varphi$, with $\kappa \in \mathbb{C}, |\kappa| = 1$, chosen to guarantee $0 \leq \alpha \leq 1$, has the stated properties. □

Using (4.2) and (4.3) one obtains

$$\begin{aligned} LG &= \alpha Lf + \beta L\varphi \\ &= \alpha\lambda f + \beta(\lambda\varphi + \xi f + \gamma h) \\ &= \lambda G + \xi\beta f + \gamma\beta h. \end{aligned} \tag{4.5}$$

For $n \geq n_w$, both $\xi \equiv \xi_n$ and $\gamma \equiv \gamma_n$ are small and G almost looks like a Dirichlet eigenfunction.

In the next sections we prove that λ, respectively, G, are good approximations of the Dirichlet eigenvalue μ_n, respectively, Dirichlet eigenfunction g.

4.2. Fourier block decomposition. Let L_{Dir} be the closed operator $L_{Dir} = -\frac{d^2}{dx^2} + V$ with domain $domL_{Dir} := \{f \in H^2[0,1] \quad | \quad f(0) = 0; \quad f(1) = 0\}$. In this section, let us fix n with $n \geq \max(n_w, 2K_8(M+1))$ (cf. Lemma 1.4 for K_8). $P_{Dir} \equiv P_{n,Dir}$ denotes the Riesz projector

$$P_{Dir} := \frac{1}{2\pi i}\int_{|z-n^2\pi^2|=M} (z - L_{Dir})^{-1} dz$$

acting on $L^2([0,1];\mathbb{C})$. Let $Q_{Dir} := Id - P_{Dir}$. Notice that

$$Q_{Dir} f \in domL_{Dir} \qquad \forall f \in domL_{Dir}, \tag{4.6}$$

$$Q_{Dir} L_{Dir} f = L_{Dir} Q_{Dir} f \qquad \forall f \in domL_{Dir}, \tag{4.7}$$

and

$$Q_{Dir} \cdot P_{Dir} = 0; \quad P_{Dir} \cdot Q_{Dir} = 0; \quad P_{Dir}^2 = P_{Dir}; \quad Q_{Dir}^2 = Q_{Dir}. \tag{4.8}$$

According to Lemma 1.5,

$$\|P_{Dir}\| \leq K_{10},$$

K_{10} being an absolute constant, and, therefore

$$\|Q_{Dir}\| \leq K_{10} + 1.$$

Notice that (cf. (1.13) and (1.16))

$$\text{Range} P_{Dir} = \{ag \mid a \in \mathbb{C}\},$$

where g is an L^2-normalized eigenfunction for the Dirichlet eigenvalue $\mu \equiv \mu_n$,

$$L_{Dir} g = \mu g; \qquad \|g\| = 1.$$

As G (cf. Lemma 4.1) is in $dom(L_{Dir})$, it admits a decomposition

$$G = P_{Dir} G + Q_{Dir} G = \kappa g + u, \tag{4.9}$$

where $u \in \text{Range}(Q_{Dir}) \subset dom(L_{Dir})$. Therefore,

$$L_{Dir} G = \kappa\mu g + L_{Dir} u \tag{4.10}$$

and

$$P_{Dir}u = 0; \quad Q_{Dir}u = u. \tag{4.11}$$

Hence, (4.7) implies that $Q_{Dir}L_{Dir}u = L_{Dir}u$, and thus,

$$L_{Dir}u \in \mathrm{Range}(Q_{Dir}). \tag{4.12}$$

On the other hand, by (4.5), $LG = \lambda G + R$, where

$$R := \xi\beta f + \gamma\beta h. \tag{4.13}$$

Thus by (4.9),

$$L_{Dir}G = \lambda\kappa g + \lambda u + P_{Dir}R + Q_{Dir}R. \tag{4.14}$$

The left sides of (4.10) and (4.14) being the same, we conclude that

$$\kappa\mu g + L_{Dir}u = \kappa\lambda g + \lambda u + P_{Dir}R + Q_{Dir}R. \tag{4.15}$$

This equation leads to the following lemma.

LEMMA 4.2. *Assume that* $3(1 + \|V\|_w)^2 \leq \frac{M}{4}$ *and* $n \geq \max(n_w, 2K_8(M + 1))$. *Then*

$$\kappa(\mu - \lambda)g = P_{Dir}R, \tag{4.16}$$

$$(L_{Dir} - \lambda)u = Q_{Dir}R. \tag{4.17}$$

Proof. Apply P_{Dir} to (4.15). In view of (4.7), (4.8), and (4.11)

$$P_{Dir}L_{Dir}u = P_{Dir}L_{Dir}Q_{Dir}u = P_{Dir}Q_{Dir}L_{Dir}u = 0.$$

Further, use that $P_{Dir}g = g$ and $P_{Dir}Q_{Dir}R = 0$ to conclude the identity (4.16). Similarly, by applying Q_{Dir} to (4.15), the second identity (4.17) is obtained. □

4.3. External equation. In this section we obtain estimates for the difference $\mu - \lambda$ between the nth Dirichlet eigenvalue $\mu \equiv \mu_n$ and the eigenvalue $\lambda \equiv \lambda_{2n}$. Recall that $G = \alpha f + \beta\varphi$ with $|\alpha|^2 + |\beta|^2 = 1$ (cf. (4.3)), $U \equiv U_n = Q_{Dir}G = G - \kappa g$ (cf. (4.9)), $L\varphi = \lambda\varphi + \xi f + \gamma h$ (cf. (4.2)), and $R = \beta(\xi f + \gamma h)$ (cf. (4.13)).

LEMMA 4.3. *Assume that* $3(1 + \|V\|_w)^2 \leq \frac{M}{4}$ *and* $n \geq \max(n_w, 2K_8(M + 1))$. *Then*

$$\|u_n\| \leq K_{11}\frac{1}{n}(|\xi_n| + 2|\gamma_n|),$$

where $K_8 > 0$ *is the absolute constant from Lemma* 1.4 *and* $K_{11} > 0$ *is the absolute constant from Lemma* 1.6.

Proof. Apply Lemma 1.6 to (4.17) to get

$$\|u\| \leq \|(\lambda - L_{Dir})^{-1}Q_{Dir}R\| \leq K_{11}\frac{1}{n}\|R\|.$$

By the definition (4.13) and $|\beta| \leq 1$,

$$\|R\| \leq |\xi| + 2|\gamma|,$$

where we used that $\|f\| = 1$ and $\|h\| \leq \sqrt{2}$ (cf. Lemma 3.3). □

From the estimate of $\|u\|$ we obtain an estimate of $\kappa \equiv \kappa_n$ in $G = \kappa g + u$ from below.

LEMMA 4.4. *Assume that* $3(1 + \|V\|_w)^2 \leq \frac{M}{4}$. *Then there exists* N_w *with* $N_w \geq \max(1 + n_w, 2K_8(M+1))$ *so that*

$$|\kappa_n| \geq \frac{1}{2} \qquad \forall n \geq N_w.$$

Proof. By (4.9), $|\kappa| = \|\kappa g\| = \|G - u\| \geq \|G\| - \|u\| = 1 - \|u\|$. By Lemma 4.3, $\|u\| \leq K_{11}\frac{1}{n}(|\xi| + 2|\gamma|)$, and by Theorem 3.5

$$|\xi_n| \leq 120(1 + n_w)^2(1 + \|V\|_w)^2 \quad \forall n \geq n_w$$

and (cf. Proposition 2.16)

$$|\gamma_n| \leq 8(1 + n_w)^{3/2}(1 + \|V\|_w)^2.$$

Thus, for $n \geq N_w$, with N_w defined by

$$N_w := 300(K_8 + K_{11})(1 + n_w)^2(1 + \|V\|_w)^2; \quad N_w \geq e, \tag{4.18}$$

it follows that $\|u\| \leq 1/2$, and thus $|\kappa| \geq 1/2$. □

4.4. Estimates for the Dirichlet eigenvalues. From the identity (4.16) we deduce an estimate for $\mu - \lambda$, using the bound for κ established in Lemma 4.4.

THEOREM 4.5. *Assume* $3(1 + \|V\|_w)^2 \leq \frac{M}{4}$. *Then, for any* $n \geq N_w$, *with* N_w *given by* (4.18),

$$|\mu_n - \lambda_n^+| \leq 2 \cdot K_{10}(|\xi_n| + 2|\gamma_n|),$$

where K_{10} *is an absolute constant given by Lemma* 1.5.

Proof. By (4.16)

$$|\kappa||\mu - \lambda| \leq \|P_{Dir}\|\|R\| \leq K_{10}\|R\| \tag{4.19}$$

and, by (4.13)

$$\|R\| \leq |\xi| + 2|\gamma|, \tag{4.20}$$

where we used that $\|f\| = 1, \|h\| \leq \sqrt{2}$, and $|\beta| \leq 1$. Combine the estimates (4.19) and (4.20) with the estimate $|\kappa| \geq \frac{1}{2}$ of Lemma 4.4 to obtain the claimed statement. □

Combined with the estimates for γ_n (Proposition 2.16) and for ξ_n (Theorem 3.5) we obtain the following theorem.

THEOREM 4.6. *Assume* $3(1 + \|V\|_w)^2 \leq \frac{M}{4}$. *Then, with* N_w *given by* (4.18),

$$\left(\sum_{n \geq N_w} w(2n)^2|\mu_n - \lambda_n^+|^2\right)^{1/2} \leq 300K_{10}13^2(1 + \|V\|_w)^6.$$

Proof. By Theorem 3.5,

$$\left(\sum_{n \geq n_w} w(2n)^2|\xi_n|^2\right)^{1/2} \leq 120(1 + n_w)^2(1 + \|V\|_w)^2.$$

By Proposition 2.16

$$\left(\sum_{n\geq n_w} w(2n)^2|\lambda_n^+-\lambda_n^-|^2\right)^{1/2} \leq 80(1+n_w)^{3/2}(1+\|V\|_w)^2.$$

Apply this to Theorem 4.5 to obtain the claimed statements. □

5. Spectrum for a special class of boundary conditions.

5.1. Special class of boundary conditions. An elementary observation (Lemma 4.1) provided us with a nonzero function G_n in the periodic or antiperiodic 2-dimensional Riesz subspace E_n (cf. (4.1)) which satisfied Dirichlet boundary conditions. If the boundary condition has such a feature, the results of section 4 can be extended. This is explained in section 5.2.

We ask the question which boundary conditions *bc*, given by two linearly independent, homogeneous equations, have the property that, for any n, the 2-dimensional subspace E_n contains a nonzero function satisfying these *bc*. Any boundary conditions *bc* for the operator $L = -\frac{d^2}{dx^2} + V$ on $[0,1]$, given by two linearly independent homogeneous equations, is a 2-dimensional subspace $\mathcal{E}$ in

$$\mathbb{C}^4 = \mathbb{C}^2 \times \mathbb{C}^2 = \{(y_0, y_0';\, y_1, y_1')\},$$

where we think of $y_0 = y(0), y_0' = y'(0), y_1 = y(1)$, and $y_1' = y'(1)$ as given by a solution $y(x) \equiv y(x,\lambda)$ of $Ly = \lambda y$. We want $\mathcal{E}$ to have a nontrivial intersection with both 2-dimensional subspaces

$$\mathcal{E}^+ := \{(y_0, y_0';\, y_1, y_1') \in \mathbb{C}^4 \mid y_0 = y_1;\, y_0' = y_1'\}$$

and

$$\mathcal{E}^- := \{(y_0, y_0';\, y_1, y_1') \in \mathbb{C}^4 \mid y_0 = -y_1;\, y_0' = -y_1'\},$$

i.e., with 2-dimensional planes of periodic and antiperiodic boundary conditions. It implies that

$$\dim(\mathcal{E} \cap \mathcal{E}^+) \geq 1; \quad \dim(\mathcal{E} \cap \mathcal{E}^-) \geq 1.$$

But

$$\mathcal{E}^+ \cap \mathcal{E}^- = \{0\},$$

which is obvious from the definition of $\mathcal{E}^+$ and $\mathcal{E}^-$. Therefore,

$$\dim(\mathcal{E} \cap \mathcal{E}^\pm) = 1 \tag{5.1}$$

and

$$\mathcal{E} \cap \mathcal{E}^+ = \{ze^+ | z \in \mathbb{C}\}; \quad e^+ := (a,b;a,b) \neq 0, \tag{5.2}$$

$$\mathcal{E} \cap \mathcal{E}^- = \{ze^- | z \in \mathbb{C}\}; \quad e^- := (c,d;-c,-d) \neq 0. \tag{5.3}$$

We conclude that

$$\mathcal{E} = \{\xi e^+ + \eta e^- | \xi, \eta \in \mathbb{C}\}.$$

It is easy to see that the orthogonal complement of $\mathcal{E}$ in $\mathbb{C}^4$ is given by

$$\mathcal{E}^\perp = \{\xi \ell_1 + \eta \ell_2 | \xi, \eta \in \mathbb{C}\},$$

where $\ell_1 = (b, -a; b, -a)$ and $\ell_2 = (d, -c; -d, c)$. Hence,

$$\mathcal{E} = \{(y_0, y_0'; y_1, y_1') \in \mathbb{C}^2 \times \mathbb{C}^2 | b(y_0 + y_1) - a(y_0' + y_1') = 0; d(y_0 - y_1) - c(y_0' - y_1') = 0\}.$$

In this way, we come, by necessity, to the following two homogeneous linear equations:

$$\begin{aligned} b(y_0 + y_1) - a(y_0' + y_1') &= 0, \\ d(y_0 - y_1) - c(y_0' - y_1') &= 0. \end{aligned} \tag{5.4}$$

They are linearly independent for any pairs $(a, b) \neq 0,\ (c, d) \neq 0$ given by (5.2) and (5.3).

We can assume without loss of generality that

$$|a|^2 + |b|^2 = 1; \quad |c|^2 + |d|^2 = 1. \tag{5.5}$$

5.2. Spectrum for L_{bc} with bc of class $\mathcal{B}$. In this section we consider only *regular boundary conditions* (see [16, section 4.8(b)] of the type (5.4). A simple verification along the definition [16, section 4.8(b), (39)] shows that the *boundary conditions* (5.4) *are regular iff*

$$ac \neq 0 \qquad \text{or} \qquad a = c = 0. \tag{5.6}$$

We denote by $\mathcal{B}$ the class of boundary conditions (5.4) which satisfy (5.5) and (5.6).

Examples. (i) Dirichlet $bc : (a, b) = (c, d) = (0, 1)$.

(ii) Neumann $bc : (a, b) = (c, d) = (1, 0)$.

(iii) More generally, if $(a, b) = e^{i\theta}(c, d)$, i.e., $\det\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = 0$, and $ac \neq 0$, let $\beta := \frac{b}{a}$. Then the boundary conditions bc (5.4) can be rewritten as

$$y'(0) = \beta y(0); \quad y'(1) = \beta y(1),$$

so bc splits and the conditions at the left and right end points of the interval $[0, 1]$ are the same.

Let us analyze $spec(L_{bc})$ for the potential $V = 0$ and boundary conditions bc from the class $\mathcal{B}$. The domain of L_{bc} is defined as

$$domL_{bc} := \{f \in H^2[0, 1] | (f_0, f_0';\ f_1, f_1') \in (5.4)\}.$$

We write, routinely,

$$\begin{gathered} -f'' = \lambda f; \qquad \lambda = \omega^2; \\ f = p \cos \omega x + q \frac{\sin \omega x}{\omega} \end{gathered}$$

and try to find all ω's such that, with this f, the linear system (5.4) has a nonzero solution $(p, q) \in \mathbb{C}^2$. This leads to the characteristic equation

$$(bd + ac\omega^2) \frac{\sin \omega}{\omega} = 0$$

or

$$(bd + ac\lambda)\frac{\sin\sqrt{\lambda}}{\sqrt{\lambda}} = 0.$$

If $a = c = 0$ (Dirichlet *bc*, cf. example (i) above), then

$$spec(L_{bc}) = \{\pi^2 k^2 | k \in \mathbb{Z}_{\geq 1}\}$$

and all eigenvalues are simple. If $ac \neq 0$,

$$spec(L_{bc}) = \{\lambda_0\} \cup \{\pi^2 k^2 | k \in \mathbb{Z}_{\geq 1}\},$$

where $\lambda_0 = -\frac{bd}{ac}$. In this case all eigenvalues are simple, except if $\lambda_0 = \pi^2 k^2$ for some $k \in \mathbb{Z}_{\geq 1}$.

Now we can claim that for any $V \in L^2[0,1]$ and $L = -\frac{d^2}{dx^2} + V$, the operator L_{bc} with *bc* from the class $\mathcal{B}$ has a discrete spectrum *spec*(L_{bc}) which consists, up to a possible additional eigenvalue ν_0, of a sequence $(\nu_n)_{n\geq 1}$ which we enumerate as in (1.5). Further, the operator L_{bc}, its resolvent and Riesz projectors have all the properties stated in Lemmas 1.4–1.6 with obvious semantic adjustments.

Property (5.1) gives a substitute for Lemma 4.1. Now we have all the tools to repeat the constructions and the proofs of section 4 for *bc* in the class $\mathcal{B}$ to get the following theorem.

Theorem 5.1. *There exist absolute constants K_{12}, K_{13} such that for any 1-periodic potential V in H_0^w and any bc in the class $\mathcal{B}$,*

$$\sum_{n\geq N} w(2n)^2 |\nu_n - \lambda_{2n}|^2 \leq K_{12}(1 + \|V\|_w)^{K_{13}},$$

where $N = K_{12}(1 + \|V\|_w)^{K_{13}}$.

Appendix. We present two lemmas used in section 1.7 concerning the convolution operation in sequence spaces. For a weight $v = (v(k))_{k\in\mathbb{Z}}$ let

$$C_v^2 := \sup_{m\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \left(\frac{v(m)}{v(k)v(m-k)}\right)^2$$

and denote by $\ell_v^2 \equiv \ell_v^2(\mathbb{Z};\mathbb{C})$ the space of sequences $(a(k))_{k\in\mathbb{Z}}$ with

$$\|a\|_v := \left(\sum_k v(k)^2 |a(k)|^2\right)^{1/2} < \infty.$$

Lemma A.1. *If $C_v < \infty$, then $\ell_v^2(\mathbb{Z};\mathbb{C})$ is a convolution algebra and*

$$\|a * b\|_v \leq C_v \|a\|_v \|b\|_v.$$

Remark. Lemma A.1 is a special case of a much more general result due to Nikolski [17].

Proof of Lemma A.1. Let $a, b \in \ell_v^2$ be with norm 1 and define $c = (c(k))_{k\in\mathbb{Z}}$ by

$$c(k) := \sum_j a(k-j)b(j).$$

We have to show that $(v(k)c(k))_{k\in\mathbb{Z}}$ is a sequence in ℓ^2. Let $\alpha(k) := v(k)|a(k)|$ and $\beta(k) := v(k)|b(k)|$ $(k \in \mathbb{Z})$.

For any sequence $\gamma(k))_{k\in\mathbb{Z}}$ in ℓ^2,

$$\begin{aligned}\sum_k v(k)|c(k)||\gamma(k)| &\leq \sum_k \sum_j |\gamma(k)| \frac{v(k)}{v(k-j)v(j)} \alpha(k-j)\beta(j) \\ &\leq \left(\sum_{i,j} |\frac{v(i+j)}{v(i)v(j)}|^2 |\gamma(i+j)|^2 \right)^{1/2} \left(\sum_{i,j} (\alpha(i)\beta(j))^2 \right)^{1/2},\end{aligned}$$

where for the last inequality we used Cauchy's inequality in $\ell^2(\mathbb{Z}\times\mathbb{Z})$. As $\sum_{i,j} \alpha(i)^2\beta(j)^2 = 1$ by assumption and

$$\sum_{i,j} \left| \frac{v(i+j)}{v(i)v(j)} \gamma(i+j) \right|^2 \leq \sum_m \sum_j \left| \frac{v(m)}{v(m-j)v(j)} \right|^2 |\gamma(m)|^2 \leq C_v^2 \sum_m |\gamma(m)|^2$$

with $(\gamma(m))_{m\in\mathbb{Z}} \in \ell^2$ arbitrary, it follows that $\|c\|_v \leq C_v$. □

For any submultiplicative weight $(w(k))_{k\in\mathbb{Z}}$ and $\alpha \geq 0$, define

$$w_\alpha(k) := \left(1 + \frac{|k|}{2}\right)^\alpha w(k).$$

Notice that w_α is again submultiplicative.

LEMMA A.2. *If* $\alpha > 1/2$, *then* $C_{w_\alpha} < \infty$.

Proof of Lemma A.2. As w is submultiplicative,

$$\frac{w_\alpha(k)}{w_\alpha(j)w_\alpha(k-j)} \leq \frac{(1+\frac{1}{2}|k|)^\alpha}{(1+\frac{1}{2}|j|)^\alpha(1+\frac{1}{2}|k-j|)^\alpha},$$

and therefore,

$$C_{w_\alpha} \leq C(\alpha),$$

where $C(\alpha) < \infty$ is a constant satisfying

$$\sum_j \left(\left(1 + \frac{|j|}{2}\right) \left(1 + |\frac{k-j}{2}|\right) \right)^{-2\alpha} \leq C(\alpha)^2 \left(1 + \frac{|k|}{2}\right)^{-2\alpha}. \tag{A.1}$$

By an elementary computation one could show that

$$C(\alpha)^2 \leq 2(1+2^{2\alpha})\frac{2\alpha+1}{2\alpha-1}. \qquad \square \tag{A.2}$$

REFERENCES

[1] J. Avron and B. Simon, *The asymptotics of the gap in the Mathieu equation*, Ann. Physics, 134 (1981), pp. 76–84.

[2] B.A. Dubrovin, I.M. Krichever, and S.P. Novikov, *Integrable systems* I, in Encyclopedia Math. Sci., Dynamical Systems IV, V.I. Arnold and S.P. Novikov, eds., Springer, Berlin, 1990, pp. 173–280.

[3] M.G. Gasymov, *Spectral analysis of a class of second-order nonselfadjoint differential operators*, Funct. Anal. Appl., 14 (1980), pp. 14–19.

[4] A. Grigis, *Estimations asymptotiques des intervalles d'instabilité pour l'équation de Hill*, Ann. Sci. École Norm. Sup. (4), 20 (1987), pp. 641–672.

[5] V. Guillemin and A. Uribe, *Hardy functions and the inverse spectral method*, Comm. Partial Differential Equations, 8 (1983), pp. 1455–1474.

[6] E. Harrell, *On the effect of the boundary condition on the eigenvalues of ordinary differential equations*, Amer. J. Math., Supplement 1981, John Hopkins University Press, Baltimore, 1981.

[7] H. Hochstadt, *Estimates on the stability intervals for Hill's equation*, Proc. Amer. Math. Soc., 14 (1963), pp. 930–932.

[8] T. Kappeler and B. Mityagin, *Gap estimates of the spectrum of Hill's equation and action variables for KdV*, Trans. Amer. Math. Soc., 351 (1999), pp. 595–617.

[9] M.V. Keldysh, *On the eigenvalues and eigenfunctions of certain classes of nonselfadjoint equations*, Dokl. Akad. Nauk, 77 (1951), pp. 11–14.

[10] M.V. Keldysh, *On the completeness of the eigenfunctions in certain classes of nonselfadjoint operators*, Uspekhi. Mat. Nauk, 27 (1971), pp. 15–41.

[11] B.M. Levitan and I.S. Sargsian, *Introduction to spectral theory*, Transl. Math. Monogr. 39, AMS, Providence, RI, 1975.

[12] W. Magnus and S. Winkler, *Hill's Equation*, Interscience, New York, 1969.

[13] V.A. Marčenko, *Sturm-Liouville Operators and Applications*, Birkhäuser, Boston, 1986.

[14] A.S. Markus, *Introduction to the spectral theory of polynomial operator pencils*, Transl. Math. Monogr. 71, AMS, Providence, 1988.

[15] H.P. McKean and E. Trubowitz, *Hill's operator and hyperelliptic function theory in the presence of infinitely many branch points*, Comm. Pure Appl. Math., 29 (1976), pp. 143–226.

[16] M.A. Neumark, *Lineare Differentialoperatoren*, Akademie-Verlag, Berlin, 1963.

[17] N. Nikolski, *Selected Problems of Weighted Approximation and Spectral Theory*, Proc. Steklov Inst. Math. 120, AMS, Providence, 1976.

[18] J. Pöschel and E. Trubowitz, *Inverse Spectral Theory*, Academic Press, New York, 1987.

[19] J.J. Sansuc and V. Tkachenko, *Spectral properties of non-selfadjoint Hill's operators with smooth potentials*, in Algebraic and Geometric Methods in Mathematical Physics, A. Boutet de Monvel and V. Marčenko, eds., Kluwer, Dordrecht, 1996, pp. 371–385.

[20] E. Trubowitz, *The inverse problem for periodic potentials*, Comm. Pure Appl. Math., 30 (1977), pp. 321–342.

SIAM J. Math. Anal.
Vol. 33, No. 1, pp. 153–171

DETERMINING CONDUCTIVITY WITH SPECIAL ANISOTROPY BY BOUNDARY MEASUREMENTS*

GIOVANNI ALESSANDRINI[†] AND ROMINA GABURRO[†]

Abstract. We prove results of uniqueness and stability at the boundary for the inverse problem of electrical impedance tomography in the presence of possibly anisotropic conductivities. We assume that the unknown conductivity has the form $A = A(x, a(x))$, where $a(x)$ is an unknown scalar function and $A(x, t)$ is a given matrix-valued function. We also deduce results of uniqueness in the interior among conductivities A obtained by piecewise analytic perturbations of the scalar term a.

Key words. inverse boundary value problems, anisotropic conductivity, singular solutions

AMS subject classifications. 35R30, 35R25

PII. S0036141000369563

1. Introduction.

In this paper we shall consider the inverse conductivity problem in an anisotropic medium. Given, in a domain $\Omega \subset \mathbb{R}^n$ (representing an electrostatic conductor), a symmetric, positive definite matrix $A = A(x)$, $x \in \Omega$ (the conductivity tensor), the Dirichlet-to-Neumann map associated to A is the operator Λ_A which, for each solution u (the electrostatic potential) of the elliptic equation

$$\operatorname{div}(A\nabla u) = 0 \quad \text{in} \quad \Omega, \tag{1.1}$$

associates to its Dirichlet data $u|_{\partial\Omega}$ (the boundary voltage) the corresponding Neumann data (the boundary current density)

$$\Lambda_A \, u|_{\partial\Omega} = A \, \nabla u \cdot \nu|_{\partial\Omega}. \tag{1.2}$$

The inverse conductivity problem then consists of determining A from the knowledge of Λ_A. While for the case when A is a priori known to be isotropic (that is, $A(x) = a(x)\, I$, where a is a scalar function) the uniqueness issue can be considered solved (see [SU], [N]), the situation is more complicated in the anisotropic case.

Since Tartar's observation [KV1] that any diffeomorphism of Ω which keeps the boundary points fixed has the property of leaving the Dirichlet-to-Neumann map unchanged, whereas A is modified, different lines of research have been pursued.

One direction has been the one of proving that the conductivity A is uniquely determined up to a change of variables in the space coordinates (see [LeU], [S], [N], [LaU]).

Another direction has been the one of assuming that the conductivity A is a priori known to depend on a restricted number of unknown spatially dependent parameters. Kohn and Vogelius [KV1] suggested the study of matrices A which are completely known with the exception of one of their eigenvalues. In [A] it is considered the case when $A(x)$ is a priori known to have the structure $A(x) = A(a(x))$, where $t \to A(t)$ is a given matrix-valued function and $a = a(x)$ is an unknown scalar function. In other words, it is assumed that at each point x the conductivity may take one value among

*Received by the editors March 22, 2000; accepted for publication (in revised form) November 27, 2000; published electronically April 24, 2001. This work was supported in part by MURST.
http://www.siam.org/journals/sima/33-1/36956.html

[†]Dipartimento di Scienze Matematiche, Università degli Studi di Trieste, Via Valerio 12, 34100 Trieste, Italy (alessang@univ.trieste.it, gaburro@mathsun1.univ.trieste.it).

a one-parameter family of admissible matrices $A(t)$ which is a priori known. In [A] results of uniqueness and stability at the boundary are proven under the additional assumption of monotonicity

$$D_t A(t) \geq \text{Const.}\, I \ > 0.$$

Lionheart [L] has proven results of uniqueness at the boundary when $A(x)$ has the structure

$$A(x) \ = \ a(x)\, A_0(x),$$

where $A_0(x)$ is given and $a \ = \ a(x)$ is an unknown scalar parameter. This structure condition may be interpreted as if at every point the anisotropic character of the conductivity were known with the exception of a scaling factor $a(x)$ which may vary from point to point.

The aim of this paper is to show that the method of singular solutions introduced in [A] enables us also to treat the case when $A(x)$ has the more general structure

$$A(x) \ = \ A(x,\, a(x)),$$

where $a(x)$ is an unknown scalar function and $A(x,\, t)$ is given and satisfies the monotonicity assumption

$$D_t A(x,\, t) \geq \text{Const.}\, I \ > 0\,.$$

We shall prove results of uniqueness and stability at the boundary which improve in various respects the results in [A] and can also be applied to the problem introduced in [L].

In Theorem 2.1 we shall prove a result of Lipschitz continuity of the boundary values of $A(x,\, a(x))$ in terms of its corresponding Dirichlet-to-Neumann map.

Theorem 2.2 gives Hölder estimates on the dependence from the Dirichlet-to-Neumann map of higher order derivatives of $A(x,\, a(x))$. This theorem is expressed in a local form. Theorem 2.3 contains the uniqueness result in the determination of $A(x,\, a(x))$ and its derivatives on the boundary. Also in this case, the result is expressed in local terms.

Theorem 2.4 gives a global uniqueness result of $A(x,\, a(x))$ among perturbations $A(x,\, b(x))$, where $a(x) - b(x)$ is piecewise analytic. The procedure under which Theorem 2.3 implies global uniqueness results in the piecewise analytic category is by now well known (see [KV2], [A], [L]); we wish to stress, however, that the present result does not require any condition of higher order differentiability on the given matrix $A(x,\, t)$; this also gives a substantial improvement to Theorem 3.4 in [L].

We conclude the paper with a discussion of the so-called one-eigenvalue-problem treated by Kohn and Vogelius [KV1]. In fact, we observe that this problem does not precisely fit the scheme of our Theorems 2.1–2.4 since, in this case, the monotonicity assumption is not satisfied. We present, however, some arguments showing how the monotonicity assumption can be relaxed in such a way that Theorems 2.1–2.4 continue to hold and, at the same time, it enables us to encompass the one-eigenvalue-problem.

The plan of the paper is as follows. In section 2 we give some basic definitions and the statements of the main Theorems 2.1–2.4. Section 3 contains the proofs of the stability results, Theorems 2.1 and 2.2. Section 4 contains the proofs of the uniqueness results, Theorems 2.3 and 2.4. Finally, section 5 contains the discussion of a generalization of the above theorems which enables us also to treat the one-eigenvalue-problem by Kohn and Vogelius.

2. Main results. In what follows we shall need the following quantitative formulation of the Lipschitz regularity of the boundary of Ω.

Definition 2.1. *Given positive numbers L, r, and h satisfying $h \geq Lr$, we say that a bounded domain $\Omega \in \mathbb{R}^n$ has a Lipschitz boundary if, for every $x^0 \in \partial\Omega$, there exists a rigid transformation of coordinates which maps x^0 into the origin such that, setting $x = (x', x_n)$, $x' \in \mathbb{R}^{n-1}$, $x_n \in \mathbb{R}$, we have*

$$\begin{aligned}&\Omega \cap \{ x = (x', x_n) \mid \ |x'| < r, |x_n| < h\} \\ &= \{ x = (x', x_n) \mid \ |x'| < r, |x_n| < h, x_n \geq f(x') \},\end{aligned}$$

where $f = f(x')$ is a Lipschitz function defined for $|x'| < r$, which satisfies

$$\begin{gathered} f(0) = 0, \\ |f(x') - f(y')| \leq L\,|x' - y'| \end{gathered}$$

for every x', $y' \in \mathbb{R}^{n-1}$, with $|x'|$, $|y'| < r$.

Let us introduce here the class of functions $A(x, t)$ which will be considered as admissible conductivities in our results.

Definition 2.2. *Given $p > n$, $E > 0$, and denoting by Sym_n the class of $n \times n$ real-valued symmetric matrices, we say that $A(\cdot, \cdot) \in \mathcal{H}$ if the following conditions hold:*

$$A \in W^{1,p}(\Omega \times [\lambda^{-1}, \lambda], Sym_n), \tag{2.1}$$

$$D_t A \in W^{1,p}(\Omega \times [\lambda^{-1}, \lambda], Sym_n), \tag{2.2}$$

$$\begin{aligned} supess_{t \in [\lambda^{-1}, \lambda]} \Big(&\| A(\cdot, t) \|_{L^p(\Omega)} + \| D_x A(\cdot, t) \|_{L^p(\Omega)} \\ &+ \| D_t A(\cdot, t) \|_{L^p(\Omega)} + \| D_t D_x A(\cdot, t) \|_{L^p(\Omega)} \Big) \leq E, \end{aligned} \tag{2.3}$$

$$\begin{aligned} \lambda^{-1}|\xi|^2 \leq A(x,t)\xi \cdot \xi \leq \lambda|\xi|^2 \quad &for\ almost\ every\ x \in \Omega, \\ &for\ every\ t \in [\lambda^{-1}, \lambda],\ \xi \in \mathbb{R}^n, \end{aligned} \tag{2.4}$$

$$\begin{aligned} D_t A(x,t)\, \xi \cdot \xi \geq E^{-1}|\xi|^2 \quad &for\ almost\ every\ x \in \Omega, \\ &for\ every\ t \in [\lambda^{-1}, \lambda],\ \xi \in \mathbb{R}^n. \end{aligned} \tag{2.5}$$

We observe that (2.4) is a condition of uniform ellipticity, whereas (2.5) is a condition of monotonicity with respect to the last variable t. Denoting by $\langle \cdot, \cdot \rangle$ the $L^2(\partial\Omega)$-pairing between $H^{\frac{1}{2}}(\partial\Omega)$ and its dual $H^{-\frac{1}{2}}(\partial\Omega)$, the Dirichlet-to-Neumann map

$$\Lambda_{A(x,a)} : H^{\frac{1}{2}}(\partial\Omega) \longrightarrow H^{-\frac{1}{2}}(\partial\Omega)$$

is defined by

$$\langle \Lambda_{A(x,a)}\, u, \phi \rangle = \int_\Omega A(x, a(x)) \nabla u(x) \cdot \nabla \phi(x)\, dx$$

for any $\phi \in H^1(\Omega)$ and for any $u \in H^1(\Omega)$ which is a weak solution to

$$\operatorname{div}(A(x,\, a(x))\, \nabla u(x)) \;=\; 0.$$

We shall denote by $\|\cdot\|_*$ the norm on the Banach space of bounded linear operators between $H^{\frac{1}{2}}(\partial\Omega)$ and $H^{-\frac{1}{2}}(\partial\Omega)$.

THEOREM 2.1 (Lipschitz stability of boundary values). *Given $p > n$, let Ω be a bounded domain with Lipschitz boundary with constants L, r, h. Let a, b satisfy*

$$\lambda^{-1} \le a(x), b(x) \le \lambda \qquad \textit{for every } x \in \Omega, \tag{2.6}$$

$$\| a \|_{W^{1,p}(\Omega)}\,,\ \| b \|_{W^{1,p}(\Omega)} \le E. \tag{2.7}$$

Let $A \in \mathcal{H}$; then we have

$$\| A(x,\, a(x)) - A(x,\, b(x)) \|_{L^\infty(\partial\Omega)} \le C \, \| \Lambda_{A(x,\, a)} - \Lambda_{A(x,\, b)} \|_* \,. \tag{2.8}$$

Here $C > 0$ is a constant which depends only on n, p, L, r, h, $diam(\Omega)$, λ, and E.

THEOREM 2.2 (Hölder stability of derivatives at the boundary). *Let a, b satisfy* (2.6), (2.7) *and let $A \in \mathcal{H}$. Given $y \in \partial\Omega$ and a neighborhood U of y in $\bar\Omega$, assume that, for some positive integer k and some α, $0 < \alpha < 1$, we have*

$$A \in C^{k,\,\alpha}(\,\bar U \times [\lambda^{-1}, \lambda]\,), \tag{2.9}$$

$$\| A \|_{C^{k,\,\alpha}(\,\bar U \times [\lambda^{-1},\,\lambda])} \le E_k, \tag{2.10}$$

$$\| a - b \|_{C^{k,\,\alpha}(\,\bar U)} \le E_k. \tag{2.11}$$

Then, for every neighborhood W of y in $\bar\Omega$ such that $\bar W \subset U$,

$$\begin{aligned} &\| D^k(A(x,a(x)) - A(x,b(x))) \|_{L^\infty(\partial\,\Omega\,\cap\,\bar W)} \\ &\le C \, \| \Lambda_{A(x,\, a)} - \Lambda_{A(x,\, b)} \|_*^{\delta_k\, \alpha}, \end{aligned} \tag{2.12}$$

where

$$\delta_k \;=\; \prod_{j\,=\,0}^{k} \frac{\alpha}{\alpha + j}. \tag{2.13}$$

Here $C > 0$ is a constant which depends only on n, p, L, r, h, $diam(\Omega)$, $dist(W \cap \partial\Omega,\, \Omega \setminus U)$, λ, E, α, k, and E_k.

THEOREM 2.3 (uniqueness at the boundary). *Let a, b satisfy* (2.6), (2.7) *and let $A \in \mathcal{H}$. Given $y \in \partial\Omega$ and a neighborhood U of y in $\bar\Omega$, assume that, for some positive integer k, we have*

$$a - b \in C^k(\bar U)\,. \tag{2.14}$$

If

$$\Lambda_{A\,(x,\, a(x))} \;=\; \Lambda_{A\,(x,\, b(x))},$$

then

$$D^j(a-b) \;=\; 0 \qquad \text{on } \partial\Omega \cap \bar{U} \qquad \text{for all } j \le k. \tag{2.15}$$

If, in addition, we have

$$A \in C^{\,k}\Big(\bar{U} \times [\lambda^{-1}, \lambda]\Big), \tag{2.16}$$

then

$$D^j(A(x\,,\,a(x))) \;=\; D^j(A(x\,,\,a(x))) \qquad \text{on } \partial\Omega \cap \bar{U} \qquad \text{for all } j \le k. \tag{2.17}$$

Theorem 2.4 (uniqueness in the interior). *Let a, b satisfy* (2.6), (2.7) *with $p = \infty$. Let $A \in \mathcal{H}$ and, in addition, $A \in W^{1,\,\infty}\big(\Omega \times [\lambda^{-1}, \lambda],\ Sym_n\big)$. Suppose that Ω can be partitioned into a finite number of Lipschitz domains, $\{A_j\}_{j\,=\,1,\ldots\,,N}$ such that $a - b$ is analytic on each $\bar{A}_j$.*

If

$$\Lambda_{A(x,\,a)} \;=\; \Lambda_{A(x,\,b)},$$

then we have

$$A(x,\, a(x)) \;=\; A(x,\, b(x)) \qquad \text{in } \Omega. \tag{2.18}$$

3. Proofs of the stability theorems. We need to introduce a unitary vector field $\tilde{\nu}$ locally defined near $\partial\Omega$ such that (i) $\tilde{\nu}$ is $C^{\,\infty}$ smooth, and (ii) $\tilde{\nu}$ is nontangential to $\partial\Omega$. To this purpose we shall make use of the following lemmas.

Lemma 3.1. *For every $x^0 \in \partial\Omega$, let $(x^{\prime},\, x_n)$ be the coordinates suited for the local representation of $\partial\Omega$ given by Definition* 2.1. *Let $x \;=\; (x^{\prime},\; f(x^{\prime}))$ be such that $|x^{\prime}| \le \rho$, where $\rho \;=\; \frac{h}{2L}$, then, picking $l \;=\; \frac{h}{2}$, we have that the truncated cone*

$$T_l(x) = \{\, z = (z^{\prime},\, z_n) \mid f(x^{\prime}) - l < z_n < f(x^{\prime}), \\ |z^{\prime} - x^{\prime}| < -L\,(f(x^{\prime}) - z_n)\}$$

has an empty intersection with Ω.

Proof. It suffices to verify that the base of $T_l(x)$ is contained in the cylinder

$$C_{r,\,h} \;=\; \{x \;=\; (x^{\prime},\, x_n)|\; |x^{\prime}| < r,\; |x_n| < h\}.$$

Hence, the verification consists of elementary calculations. □

Lemma 3.2. *There exists a finite number of points $x^1, \ldots\,, x^k \in \partial\Omega$ and rotations $R_l : \mathbb{R}^n \longrightarrow \mathbb{R}^n$, $l = 1, \ldots\,, k$, such that* (i) *the open cylinders $V_l = x^{\,l} + R_{\,l} C_{\rho,\,h}$ cover $\partial\Omega$,* (ii) *the axis of each cylinder V_l coincides with the nth coordinate axis for the local representation of $\partial\Omega$ near $x^{\,l}$ given by Definition* 2.1, *and* (iii) *for every $x \in \partial\Omega \cap V_l$, $l \;=\; 1, \ldots\,, k$, the truncated cone $x + R_{\,l} T_{\,l}(0)$ does not intersect Ω.*

Proof. The proof follows easily from the previous lemma and the compactness of $\partial\Omega$. □

Lemma 3.3. *For any $x^0 \in \partial\Omega$, let V_l be the cylinder introduced in the previous lemma such that $x^0 \in V_l$. Setting $\tilde{\nu}$ as the nth coordinate unit vector along the axis of V_l which points to the exterior of Ω, we have that the point*

$$z_\sigma \;=\; x^0 + \sigma\tilde{\nu} \tag{3.1}$$

satisfies

$$C\sigma \le d(z_\sigma, \partial\Omega) \le \sigma \qquad \textit{for every } \sigma,\ 0 \le \sigma \le \sigma^0\,, \tag{3.2}$$

where σ^0 *and* C *depend only on* L, r, h.

Proof. The proof is an elementary consequence of Definition 2.1 and the previous lemmas. □

Let us recall some results about singular solutions of elliptic equations. We consider elliptic operators

$$L\ =\ \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial}{\partial x_j}\right) \qquad \text{in } B_R\ =\ \left\{x \in \mathbb{R}^n |\, |x| < R\right\}, \tag{3.3}$$

where the coefficient matrix $(a_{ij}(x))$ is symmetric and satisfies

$$\lambda^{-1}|\,\xi|^2 \le a_{ij}(x)\xi_i\,\xi_j \le \lambda|\,\xi|^2 \qquad \text{for every } x,\ \xi,\ x \in B_R,\ \xi \in \mathbb{R}^n, \tag{3.4}$$

and also

$$\|\, a_{ij}\,\|_{\,W^{1,\,p}(B_R)} \le E, \qquad i,\ j\ =\ 1, \dots, n, \tag{3.5}$$

where $p > n$ and λ, E are positive constants.

Theorem 3.4 (singular solutions). *Let* L *satisfy* (3.3)–(3.5). *For every spherical harmonic* S_m *of degree* $m\ =\ 0,\ 1,\ 2, \dots$, *there exists* $u \in W^{2,\,p}_{loc}(B_R \setminus \{0\})$ *such that*

$$Lu\ =\ 0 \qquad \textit{in } B_R \setminus \{\,0\}, \tag{3.6}$$

and furthermore

$$u(x) = log\,|\, Jx\,|\, S_0\left(\frac{Jx}{|\,Jx\,|}\right) + w(x), \qquad when\, n\ =\ 2\, and\, m = 0, \tag{3.7}$$

$$u(x) = |\,Jx\,|^{\,2-n-m}\, S_m\left(\frac{Jx}{|\,Jx\,|}\right) + w(x) \qquad otherwise, \tag{3.8}$$

where J *is the positive definite symmetric matrix such that* $J\ =\ \sqrt{(a_{ij}(0))^{-1}}$ *and* w *satisfies*

$$|\,w(x)| + |\,x\,|\,|Dw(x)| \le C\,|\,x\,|^{\,2-n-m+\alpha} \quad in \quad B_R \setminus \{\,0\,\}, \tag{3.9}$$

$$\left(\int_{r<|x|<2r} |D^2w|^p\right)^{\frac{1}{p}} \le C\, r^{-n-m+\alpha+\frac{n}{p}} \quad \textit{for every } r,\ 0 < r < R/2. \tag{3.10}$$

Here α *is any number such that* $0 < \alpha < 1 - \frac{n}{p}$, *and* C *is a constant depending only on* α, n, p, R, λ, *and* E.

Proof. See Theorem 1.1 in [A]. □

We shall also need the following lemma.

Lemma 3.5. *Let the hypotheses of Theorem* 3.4 *be satisfied. For every* $m\ =\ 0,\ 1,\ 2, \dots$, *there exists a spherical harmonic* S_m *of degree* m *such that the solution* u *on* (3.6) *given by Theorem* 3.4 *also satisfies*

$$|Du(x)| > |x|^{\,1-(n+m)} \tag{3.11}$$

for every x, $0 < |x| < r_0$, where $r_0 > 0$ depends only on λ, E, p, m, and R.

Proof. The proof of this lemma can be obtained along the same lines as the proof of [A, Lemma 3.1]. □

LEMMA 3.6. *Let Ω be a domain in $\mathbb{R}^n$ $(n \geq 2)$ with Lipschitz boundary $\partial\Omega$. Let $A \in \mathcal{H}$ and let a be a function satisfying conditions* (2.6), (2.7). *Then we have*

$$A(\cdot,\, a(\cdot)) \in W^{1,\,p}(\Omega,\, Sym_n), \tag{3.12}$$

and furthermore,

$$\| A(\cdot,\, a(\cdot)) \|_{W^{1,\,p}(\Omega)} \leq CE(1+ \| a \|_{W^{1,\,p}(\Omega)}), \tag{3.13}$$

where C is a positive constant depending only on λ, Ω, n, and p.

Proof. We observe that the two functions

$$\begin{aligned} t &\longrightarrow A(x,\, t), \\ t &\longrightarrow D_x A(x,\, t) \end{aligned}$$

are absolutely continuous functions for almost every $x \in \Omega$ (see [M, Lemma 3.1.1]). Then the following identities hold:

$$A(x,\, a(x)) = A(x,\, \lambda) - \int_{a(x)}^{\lambda} D_t A(x,\, t)\, dt\,, \tag{3.14}$$

$$D_x A(x,\, t)|_{t\,=\,a(x)} = D_x A(x,\, \lambda) - \int_{a(x)}^{\lambda} D_t D_x A(x,\, t)\, dt \tag{3.15}$$

for almost every $x \in \Omega$.

We obtain

$$\| A(\cdot,\, a(\cdot)) \|_{L^p(\Omega)} \leq \lambda\, E. \tag{3.16}$$

Similarly, by (3.15) and by the Sobolev inequality

$$\| D_t A(\cdot,\, t) \|_{L^\infty(\Omega)} \leq C\Big(\| D_t A(\cdot,\, t) \|_{L^p(\Omega)} + \| D_x D_t A(\cdot,\, t) \|_{L^p(\Omega)} \Big),$$

we deduce

$$\| D_x A(\cdot,\, a(\cdot)) \|_{L^p(\Omega)} \leq \lambda\, E + CE \| D_x a \|_{L^p(\Omega)}\,. \tag{3.17}$$

By (3.16), (3.17), the proof is completed. □

Proof of Theorem 2.1. We start from the identity (see, for instance, [A])

$$\int_\Omega (A(x,\, a) - A(x,\, b)) Du \cdot Dv \;=\; \langle (\Lambda_{A(x,\,a)} - \Lambda_{A(x,\,b)})\, u,\, v \rangle, \tag{3.18}$$

where u, v are two arbitrary solutions to

$$\operatorname{div}(A(x,\, a)\operatorname{grad} u) \;= 0\,, \qquad \operatorname{div}(A(x,\, b)\operatorname{grad} v) \;=\; 0\,,$$

respectively. Let $x^0 \in \partial\Omega$ be such that

$$|(a-b)(x^0)| \;= \| a-b \|_{L^\infty(\,\partial\Omega)},$$

and set, for convenience, $(a-b)(x^0) > 0$. Let V_l be one of the neighborhoods, selected in Lemmas 3.2 and 3.3, such that $x^0 \in V_l$. We let $\tilde{\nu}$ be defined according to Lemma 3.3. Let us consider $z_\sigma = x^0 + \sigma\tilde{\nu}$, where σ satisfies $0 < \sigma \leq \min\{\sigma_0, \frac{r_0}{2}\}$, where σ_0 is the number fixed in Lemma 3.3, and r_0 is the number appearing in Lemma 3.5 when the integer m is chosen to be equal to zero. We observe that we can continue $a(x)$, $b(x)$ to $B_R(z_\sigma)$ in such a way that $A(x,\, a(x))$ and $A(x,\, b(x))$ continue to satisfy uniform bounds of ellipticity and on the $W^{1,\,p}$-norm. We now consider the ball $B_\rho(z_\sigma)$, with $\rho = r_0$ and fix the two solutions $u,\, v \in W^{2,\,p}(\Omega)$ found in Theorem 3.4 having a Green function type of singularity at z_σ, that is, $m = 0$ and

$$u(x) = |J_a(x-z_\sigma)|^{\,2-n} + O(|x-z_\sigma|^{\,2-n+\alpha}), \tag{3.19}$$

$$v(x) = |J_b(x-z_\sigma)|^{\,2-n} + O(|x-z_\sigma|^{\,2-n+\alpha}), \tag{3.20}$$

where $J_a = \sqrt{A(z_\sigma,\, a(z_\sigma))^{-1}}$, $J_b = \sqrt{A(z_\sigma,\, b(z_\sigma))^{-1}}$.

Applying (3.18) to the two solutions u, v above, we obtain

$$\begin{aligned} &\| \Lambda_{A(x,\,a)} - \Lambda_{A(x,\,b)} \|_* \| u \|_{H^{\frac12}(\partial\Omega)} \| v \|_{H^{\frac12}(\partial\Omega)} \\ &\geq \left| \int_{B_\rho(z_\sigma)\,\cap\,\Omega} (A(x,\, a(x)) - A(x,\, b(x)))\, Du \cdot Dv \right| \\ &- \left| \int_{\Omega\,\setminus\,B_\rho(z_\sigma)} (A(x,\, a(x)) - A(x,\, b(x)))\, Du \cdot Dv \right|. \end{aligned} \tag{3.21}$$

Then using (3.9) we end up with

$$\begin{aligned} &\left| \int_{B_\rho(z_\sigma)\,\cap\,\Omega} \frac{(A(x,\,a) - A(x,\,b))J_a^2(x-z_\sigma)\cdot J_b^2(x-z_\sigma)}{|J_a(x-z_\sigma)|^{\,n}|J_b(x-z_\sigma)|^{\,n}} \right| \\ &\leq C \int_{B_\rho(z_\sigma)\,\cap\,\Omega} |x-z_\sigma|^{\,2-2n+\alpha} \\ &+ \int_{\Omega\,\setminus\,B_\rho(z_\sigma)} |A(x,\,a) - A(x,\,b)|\, |x-z_\sigma|^{\,2-2n} \\ &+ \| \Lambda_{A(x,\,a)} - \Lambda_{A(x,\,b)} \|_* \| u \|_{H^{\frac12}(\partial\,\Omega)} \| v \|_{H^{\frac12}(\partial\,\Omega)}\ . \end{aligned} \tag{3.22}$$

We recall that, by Lemma 3.6 and by our assumptions, $A(x,\, a(x)) \in W^{1,\,p}(\Omega)$ with $p > n$; hence $A(x, a(x))$ is Hölder continuous with exponent $\beta = 1 - \frac{n}{p}$ in $\bar{\Omega}$. Therefore,

$$A(x,\, a(x)) - A(x,\, b(x)) \;=\; A(x^0,\, a(x^0)) - A(x^0,\, b(x^0)) + O(|x-x^0|^{\beta}).$$

We obtain

$$\begin{aligned} &\int_{B_\rho(z_\sigma)\,\cap\,\Omega} \frac{J_b^{\,2}(A(x^0,\, a(x^0)) - A(x^0,\, b(x^0)))J_a^{\,2}(x-z_\sigma)\cdot(x-z_\sigma)}{|J_a(x-z_\sigma)|^{\,n}|J_b(x-z_\sigma)|^{\,n}} \\ &\leq C \int_{B_\rho(z_\sigma)\,\cap\,\Omega} |x-z_\sigma|^{\,2-2n}|x-x^0|^{\,\beta} \\ &+ C \int_{B_\rho(z_\sigma)\,\cap\,\Omega} |x-z_\sigma|^{\,2-2n+\alpha} \\ &+ \int_{\Omega\,\setminus\,B_\rho(z_\sigma)} |A(x,\,a) - A(x,\,b)|\, |x-z_\sigma|^{\,2-2n} \\ &+ \| \Lambda_{A(x,\,a)} - \Lambda_{A(x,\,b)} \|_* \| u \|_{H^{\frac12}(\partial\,\Omega)} \| v \|_{H^{\frac12}(\partial\,\Omega)}\ . \end{aligned} \tag{3.23}$$

We now consider the quantity $J_b^2(A(x^0,\, a)-A(x^0,\, b))J_a^2(x-z_\sigma)\cdot(x-z_\sigma)$, which appears on the left-hand side of (3.23). Recalling that $J_a^2 = A(z_\sigma,\, a(z_\sigma))^{-1}$, we have

$$|J_a^2 - A(x^0,\, a(x^0))^{-1}| \le C|z_\sigma - x^0|^{\,\beta} \le C\sigma^{\,\beta}$$

and likewise, $|J_b^2 - A(x^0,\, b(x^0))^{-1}| \le C\sigma^{\,\beta}$. Therefore,

$$\begin{aligned}
&J_b^2(A(x^0,\, a)-A(x^0,\, b))\, J_a^2(x-z_\sigma)\cdot(x-z_\sigma)\\
&\ge (A(x^0,\, b)^{-1} - A(x^0,\, a)^{-1})(x-z_\sigma)\cdot(x-z_\sigma)\\
&-C\sigma^{\beta}(a(x^0)-b(x^0))|x-z_\sigma|^2.
\end{aligned}$$

Using the ellipticity assumption (2.4) and the monotonicity assumption (2.5), we compute

$$\begin{aligned}
&(A(x^0,\, b)^{-1} - A(x^0,\, a)^{-1})(x-z_\sigma)\cdot(x-z_\sigma)\\
&= \left(\int_{a(x^0)}^{b(x^0)} D_t(A(x^0,\, t))^{-1}\, dt\right)(x-z_\sigma)\cdot(x-z_\sigma)\\
&= \left(\int_{a(x^0)}^{b(x^0)} -A^{-1}(x^0,\, t)\, D_tA(x^0,\, t)\, A^{-1}(x^0,\, t)\, dt\right)(x-z_\sigma)\cdot(x-z_\sigma)\\
&\ge \int_{b(x^0)}^{a(x^0)} E^{-2}\lambda^{-2}\, |x-z_\sigma|^2\, dt.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&J_b^2(A(x^0,\, a)-A(x^0, b))\, J_a^2(x-z_\sigma)\cdot(x-z_\sigma)\\
&\ge (E^{-2}\lambda^{-2} - C\sigma^{\,\beta})(a(x^0)-b(x^0))\, C\, |x-z_\sigma|^2
\end{aligned}$$

and, choosing

$$\sigma \le \Big(\frac{1}{2C}E^{-2}\,\lambda^{-2}\Big)^{\frac{1}{\beta}},$$

we obtain

$$\begin{aligned}
&J_b^2(A(x^0,\, a)-A(x^0,\, b))\, J_a^2(x-z_\sigma)\cdot(x-z_\sigma)\\
&\ge C\,(a(x^0)-b(x^0))\,|x-z_\sigma|^2.
\end{aligned}\tag{3.24}$$

By applying (3.24) to (3.23),

$$\begin{aligned}
&\|\, a-b\,\|_{L^{\,\infty}(\partial\Omega)} \int_{B_\rho(z_\sigma)\,\cap\,\Omega} |x-z_\sigma|^{\,2-2n}\\
&\le C\Big\{ \int_{B_\rho(z_\sigma)\,\cap\,\Omega} |x-z_\sigma|^{\,2-2n}|x-x^0|^{\,\beta}\\
&\quad+ \int_{B_\rho(z_\sigma)\,\cap\,\Omega} |x-z_\sigma|^{\,2-2n+\alpha}\\
&\quad+ \int_{\Omega\,\setminus\, B_\rho(z_\sigma)} |A(x,\, a)-A(x,\, b)|\,|x-z_\sigma|^{\,2-2n}\\
&\quad+ \|\,\Lambda_{A(x,\,a)} - \Lambda_{A(x,\,b)}\,\|_*\|\, u\,\|_{H^{\frac{1}{2}}(\partial\,\Omega)}\|\, v\,\|_{H^{\frac{1}{2}}(\partial\,\Omega)}\Big\}.
\end{aligned}$$

By estimating the integrals appearing above and the $H^{\frac{1}{2}}(\partial\Omega)$ norms of u and v, we obtain

$$\| a-b \|_{L^\infty(\partial\,\Omega)}\, \sigma^{2-n} \le C\Big\{\sigma^{2-n+\beta}+\sigma^{2-n+\alpha}+C + \| \Lambda_{A(x,\,a)}-\Lambda_{A(x,\,b)} \|_*\, \sigma^{2-n}\Big\};$$

see [A, Proof of Theorem 1.2] for details. Therefore,

$$\| a-b \|_{L^\infty(\partial\,\Omega)} \le C\Big\{\omega(\sigma)+ \| \Lambda_{A(x,\,a)}-\Lambda_{A(x,\,b)} \|_* \Big\}, \tag{3.25}$$

where $\omega(\sigma)\to 0$ as $\sigma \to 0$. From (3.25) we obtain the following estimate:

$$\| a-b \|_{L^\infty(\partial\,\Omega)} \le C \,\| \Lambda_{A(x,\,a)}-\Lambda_{A(x,\,b)} \|_* \,. \tag{3.26}$$

Recalling that, for almost every $x\in\Omega$, the function

$$t \longrightarrow A(x,\,t)$$

is absolutely continuous on $[\lambda^{-1},\,\lambda]$, we may write

$$\begin{aligned} |A(x,\,a(x))-A(x,\,b(x))| &= \left| \int_{b(x)}^{a(x)} D_tA(x,\,t)\,dt \right| \\ &\le \int_{b(x)}^{a(x)} \mathrm{Sup}_{\,t,\,x}|\,D_tA(x,\,t)\,|\,dt \\ &\le C\,|\,(a(x)-b(x))\,| \end{aligned}$$

for every $x\in\bar\Omega$. Taking the L^∞-norm on both sides, we obtain

$$\| A(x,\,a)-A(x,\,b) \|_{L^\infty(\partial\,\Omega)} \le C\ \| a-b \|_{L^\infty(\partial\,\Omega)} \,. \tag{3.27}$$

By combining (3.26) and (3.27) we obtain (2.8). □

Proof of Theorem 2.2. Possibly reducing the values of h, r in Definition 2.1, it suffices to consider the case when $U = V_l\cap\bar\Omega$, where V_l is one of the cylinders found in Lemma 3.2 and $W = \frac{1}{2}\,V_l\cap\bar\Omega$, where $\frac{1}{2}\,V_l$ is the cylinder having the same center as V_l and half sizes compared to V_l. First, we shall prove

$$\left\| \frac{\partial^j}{\partial\tilde\nu^j}(a-b) \right\|_{L^\infty(\partial\Omega\,\cap\,\bar W)} \le C\,\| \Lambda_{A(x,\,a)}-\Lambda_{A(x,\,b)} \|_*^{\delta_j} \quad \text{for every } j\le k, \tag{3.28}$$

where δ_j is given by (2.13) and $\tilde\nu$ is the unit vector introduced in Lemma 3.3. We proceed by induction on k. Using (3.26) in the proof of Theorem 2.1, we have that (3.28) is satisfied when $k = 0$. Let us assume that (3.28) holds when $j = k-1$, and let us prove it for $j = k$.

Let m be a positive integer. Let $x^0\in\partial\Omega\cap W$ be such that (up to exchanging the roles of a and b)

$$(-1)^k\frac{\partial^k}{\partial\tilde\nu^k}(a-b)(x^0) \;=\; \left\| \frac{\partial^k}{\partial\tilde\nu^k}(a-b) \right\|_{L^\infty\,(\partial\,\Omega\,\cap\,\bar W)} . \tag{3.29}$$

Let $z_\sigma = x^0 + \sigma\tilde{\nu}$, with $\sigma \le \min\{\sigma_0, \frac{\rho}{2}\}$, where σ_0 is the number appearing in Lemma 3.3 and $\rho = \min\{r_0, \frac{h}{4L}\}$, where r_0 is the number chosen in Lemma 3.5 in dependence of m. We consider the ball $B_\rho(z_\sigma)$; we have that $B_\rho(z_\sigma) \cap \Omega$ will be nonempty and also

$$B_\rho(z_\sigma) \cap \bar{\Omega} \subset U. \tag{3.30}$$

As we did in Theorem 2.1, we can continue $A(x, a(x))$, $A(x, b(x))$ outside of Ω. Consequently, let u, v be the solutions obtained in Theorem 3.4 having a singularity at z_σ and corresponding to the spherical harmonic S_m indicated in Lemma 3.5. We apply (3.18) to two such solutions. We now use the property

$$A(x, t) \in C^1(\bar{U} \times [\lambda^{-1}, \lambda])$$

and from the Lagrange theorem, for every $x \in \bar{U}$ there exists $t(x)$, $0 < t(x) < 1$, such that

$$\begin{aligned}\Big(A(x, a(x)) - A(x, b(x))\Big) Du \cdot Dv = {} & (a(x) - b(x)) \\ & \cdot D_t A(x, t)|_{t = c(x)} Du \cdot Dv,\end{aligned}$$

where $c(x) = a(x) + t(x)(b(x) - a(x))$, hence

$$\begin{aligned}&\| \Lambda_{A(x, a)} - \Lambda_{A(x, b)} \|_* \| u \|_{H^{\frac{1}{2}}(\partial\Omega)} \| v \|_{H^{\frac{1}{2}}(\partial\Omega)} \\ &\ge \int_\Omega (A(x, a) - A(x, b)) Du \cdot Dv \\ &\ge \int_{B_\rho(z_\sigma) \cap \Omega} (a - b)\Big(D_t A(x, t)|_{t = c}\Big) Du \cdot Dv \\ &- \int_{\Omega \setminus B_\rho(z_\sigma)} |A(x, a) - A(x, b)|\, |Du|\, |Dv|.\end{aligned}$$

From the formulas (3.8)–(3.9) we have

$$\begin{aligned}|Du - Dv| \le {} & C(|x - z_\sigma|^{1-n-m} \,|a(z_\sigma) - b(z_\sigma)| \\ & + |x - z_\sigma|^{1-n-m+\alpha}) \\ \le {} & C\Big(|x - z_\sigma|^{1-n-m} |a(x^0) - b(x^0)| \\ & + |x - z_\sigma|^{1-n-m} \sigma^\beta + |x - z_\sigma|^{1-n-m+\alpha}\Big),\end{aligned}$$

and since $|x - z_\sigma| \ge C\sigma$, for every $x \in B_\rho(z_\sigma)$ and $\alpha < \beta$,

$$\begin{aligned}|Du - Dv| \le {} & C(|x - z_\sigma|^{1-n-m} |\, a(x^0) - b(x^0)| \\ & + |x - z_\sigma|^{1-n-m+\alpha}).\end{aligned} \tag{3.31}$$

Let us compute

$$\begin{aligned}& D_t A(x, t)|_{t = c(x)} Du \cdot Dv \\ &= D_t A(x, t)|_{t = c(x)} Du \cdot Du + D_t A(x, t)|_{t = c(x)} Du \cdot (Dv - Du) \\ &\ge C\, |Du|^2 \\ &- C\, |Du| \Big\{ |x - z_\sigma|^{1-n-m} |\, a(x^0) - b(x^0)| + |x - z_\sigma|^{1-n-m+\alpha} \Big\}.\end{aligned}$$

Therefore,

$$\begin{aligned} &D_tA(x,\,t)|_{\,t\,=\,c\,(x)}Du\cdot Dv\\ &\geq\ C\,|x-z_\sigma|^{\,2-2(n+m)}-\ C|x-z_\sigma|^{\,2-2(n+m)}|\,a(x^0)-b\,(x^0)|\\ &-C\,|x-z_\sigma|^{\,2-2(n+m)}\Big(1-|\,a(x^0)-b\,(x^0)|-|x-z_\sigma|^{\,\alpha}\Big) \end{aligned}\tag{3.32}$$

for almost every $x\in B_\rho(z_\sigma)\cap\Omega$.

Recalling that (3.28) holds for $j\,=\,0$, we obtain

$$\begin{aligned} &D_tA(x,\,t)|_{\,t\,=\,c\,(x)}Du\cdot Dv\\ &\geq C|x-z_\sigma|^{\,2-2(n+m)}\Big(1-C\ \parallel\Lambda_{A(x,\,a)}-\Lambda_{A(x,\,b)}\parallel_*-|x-z_\sigma|^{\,\alpha}\Big) \end{aligned}$$

for almost every $x\in B_\rho(z_\sigma)\cap\Omega$. Let us observe that, without loss of generality, we can assume

$$\parallel\Lambda_{A(x,\,a)}-\Lambda_{A(x,\,b)}\parallel_*\leq\frac{1}{2C};\tag{3.33}$$

in fact, if we had the opposite inequality we would trivially obtain

$$\begin{aligned} \parallel D^k(a-b)\parallel_{L^{\,\infty}(\partial\,\Omega)}\ &\leq\ E_k\\ &\leq E_k(2C)^{\,\delta_k}\parallel\Lambda_{A(x,\,a)}-\Lambda_{A(x,\,b)}\parallel_*^{\delta_k}, \end{aligned}$$

which would prove the induction step. Hence, using (3.33), we have

$$D_{\,t}A(x,\,t)|_{\,t\,=\,c\,(x)}\geq C|x-z_\sigma|^{\,2-2(n+m)}\Big(\frac{1}{2}-|x-z_\sigma|^{\,\alpha}\Big)\tag{3.34}$$

for almost every $x\in B_\rho(z_\sigma)\cap\Omega$.

Possibly choosing a smaller value of ρ, we may assume that

$$|x-z_\sigma|^{\,\alpha}<\frac{1}{4}\quad\text{for every } x\in B_\rho(z_\sigma),$$

and therefore,

$$D_{\,t}A(x,\,t)|_{\,t\,=\,c\,(x)}Du\cdot Dv\geq C\,|x-z_\sigma|^{\,2-2(n+m)}\tag{3.35}$$

for almost every $x\in B_\rho(z_\sigma)\cap\Omega$.

Note that every $x\in U$ can be uniquely represented as

$$x\ =\ y-s\tilde{\nu},\tag{3.36}$$

where $y\in\partial\Omega$ and $0\leq s\leq\sigma_0$, where $0<\sigma_0<h-L\,r$.

Hence, by Taylor's formula we have

$$\begin{aligned} \left\|\frac{\partial^{\,k}}{\partial\tilde{\nu}^k}(a-b)\right\|_{L^{\,\infty}(\partial\,\Omega\,\cap\,\bar{W})}s^k\ &\leq\ k\,!\,(a-b)(x)\\ &+C\Bigg\{\sum_{j\,=\,0}^{k-1}\left\|\frac{\partial^j}{\partial\tilde{\nu}^j}(a-b)\right\|_{L^{\,\infty}(\partial\,\Omega\,\cap\,\bar{W})}s^j\\ &+\,s^k|x-x^0|^{\,\alpha}\Bigg\}. \end{aligned}\tag{3.37}$$

We obtain

$$\begin{aligned}
&\| \Lambda_{A(x,\,a)} - \Lambda_{A(x,\,b)} \|_* \| u \|_{H^{\frac{1}{2}}(\partial\,\Omega)} \| v \|_{H^{\frac{1}{2}}(\partial\,\Omega)} \\
&\geq \int_{B_\rho(z_\sigma)\,\cap\,\Omega} (a-b) D_t A(x,\,t)|_{\,t\,=\,c} Du \cdot Dv \\
&\quad - \int_{\Omega\,\setminus\,B_\rho(z_\sigma)} |A(x,\,a) - A(x,\,b)|\,|x - z_\sigma|^{\,2-2(n+m)} \\
&\geq \left\| \frac{\partial^{\,k}}{\partial\tilde{\nu}^{\,k}}(a-b) \right\|_{L^{\,\infty}(\partial\,\Omega)} \int_{B_\rho(z_\sigma)\,\cap\,\Omega} (d(x,\,\partial\Omega))^{\,k} D_t A(x,\,t)|_{\,t\,=\,c} Du \cdot Dv \\
&\quad - \sum_{j\,=\,0}^{k-1} \left\| \frac{\partial^{\,j}}{\partial\tilde{\nu}^{\,j}}(a-b) \right\|_{L^{\,\infty}(\partial\,\Omega)} \int_{B_\rho(z_\sigma)\,\cap\,\Omega} (d(x,\,\partial\Omega))^{\,j} D_t A(x,\,t)|_{\,t\,=\,c} Du \cdot Dv \\
&\quad - \int_{B_\rho(z_\sigma)\,\cap\,\Omega} (d(x,\,\partial\Omega))^{\,k} |x - x^0|^{\,\alpha} D_t A(x,\,t)|_{\,t\,=\,c} Du \cdot Dv \\
&\quad - \int_{\Omega\,\setminus\,B_\rho(z_\sigma)} |A(x,\,a) - A(x,\,b)|\,|x - z_\sigma|^{\,2-2(n+m)}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left\| \frac{\partial^{\,k}}{\partial\tilde{\nu}^{k}}(a-b) \right\|_{L^{\,\infty}(\partial\,\Omega)} \int_{B_\rho(z_\sigma)\,\cap\,\Omega} |x - z_\sigma|^{\,2-2(n+m)} (d(x,\,\partial\Omega))^{\,k} \\
&\leq C \Bigg\{ \int_{\Omega\,\setminus\,B_\rho(z_\sigma)} |\,a(x) - b(x)|\,|x - z_\sigma|^{\,2-2(n+m)} \\
&\qquad + \int_{B_\rho(z_\sigma)\,\cap\,\Omega} |x - z_\sigma|^{\,2-2(n+m)} |x - z_\sigma|^{\,2-2(n+m)} |x - x^0|^{\,\alpha} (d(x,\,\partial\Omega))^{\,k} \\
&\qquad + \sum_{j\,=\,0}^{k-1} \int_{B_\rho(z_\sigma)\,\cap\,\Omega} |x - z_\sigma|^{\,2-2(n+m)} (d(x,\,\partial\Omega))^{\,j} \left\| \frac{\partial^{\,j}}{\partial\tilde{\nu}^{\,j}}(a-b) \right\|_{L^{\,\infty}(\partial\,\Omega)} \\
&\qquad + \| \Lambda_{A(x,\,a)} - \Lambda_{A(x,\,b)} \|_* \| u \|_{H^{\frac{1}{2}}(\partial\,\Omega)} \| v \|_{H^{\frac{1}{2}}(\partial\,\Omega)} \Bigg\}.
\end{aligned}$$

Choosing m sufficiently large, depending only on k, estimating the integrals in the above formula (see [A, Proof of Theorem 1.2]), and recalling the induction hypothesis, we obtain

$$\left\| \frac{\partial^{\,k}}{\partial\tilde{\nu}^{\,k}}(a-b) \right\|_{L^{\,\infty}(\partial\,\Omega\,\cap\,\bar{W})} \leq C \left\{ \| \Lambda_{A(x,\,a)} - \Lambda_{A(x,\,b)} \|_*^{\,\delta^{\,k-1}} \sigma^{-k} + \sigma^{\,\alpha} \right\}$$

for every $0 < \sigma \leq \sigma_0$.

By optimizing the choice of σ, we obtain (3.28) for $j = k$. Let us now recall the interpolation inequality

$$\begin{aligned}
\| Df \|_{L^{\,\infty}(\partial\,\Omega\,\cap\,\bar{U})} \leq C \Bigg\{ &\left\| \frac{\partial}{\partial\tilde{\nu}} f \right\|_{L^{\,\infty}(\partial\,\Omega)} \\
&+ \| f \|_{L^{\,\infty}(\partial\,\Omega\,\cap\,\bar{U})}^{\frac{\alpha}{1-\alpha}} \| f \|_{C^{\,1+\alpha}(\bar{U})}^{\frac{1}{1+\alpha}} \Bigg\}
\end{aligned} \tag{3.38}$$

for every $f \in C^{1,\alpha}(\bar{\Omega})$ (see, for instance, (3.17) in [A, Lemma 3.2]). From (3.28) and an iterated use of (3.38), we obtain

$$\| D^k(a-b) \|_{L^\infty(\partial\Omega \cap \bar{W})} \leq C \| \Lambda_{A(x,a)} - \Lambda_{A(x,b)} \|_*^{\delta_k} . \tag{3.39}$$

Finally, we observe that, by an elementary induction argument, for every multi-index β, $|\beta| \leq k$ the following identity holds:

$$\begin{aligned} D^\beta A(x, a(x)) \; = \sum_{\gamma+\delta \leq \beta} & P_{\gamma\delta}(a(x), \dots, D^{|\delta|}a(x)) \\ & \cdot D_x^\gamma D_t^{|\delta|} A(x, a(x)), \end{aligned} \tag{3.40}$$

where $P_{\gamma\delta}$ is a polynomial. Hence, recalling hypothesis (2.9), which has not been used yet, we obtain

$$\| D^k(A(x, a(x)) - A(x, b(x))) \|_{L^\infty(\partial\Omega \cap \bar{W})} \leq C \| a - b \|^\alpha_{C^k(\partial\Omega \cap \bar{W})} , \tag{3.41}$$

which in combination with (3.39) readily implies (2.12). □

4. Proofs of the uniqueness theorems.

Proof of Theorem 2.3. Let us observe that it suffices to prove (2.15) and (2.17) on $\partial\Omega \cap \bar{W}$, where W is an arbitrary open subset of $\bar{\Omega}$ such that $\bar{W} \subset U$. Therefore, similarly to what we did in the proof of Theorem 2.2, we can reduce ourselves to the case when $U = V_l \cap \bar{\Omega}$, $W = \frac{1}{2}V_l \cap \bar{\Omega}$, where V_l is one of the cylinders found in Lemma 3.2. Let $\tilde{\nu}$ be the unit vector introduced in Lemma 3.3, suited for the neighborhood U. As a first step, let us prove

$$\frac{\partial^j}{\partial \tilde{\nu}^j}(a-b) \; = \; 0 \qquad \text{on } \partial\Omega \cap \bar{W} \quad \text{for every } j \leq k, \tag{4.1}$$

by induction on k. When $k = 0$, (4.1) is a consequence of Theorem 2.1 (see also (3.26)). Let us assume that (4.1) holds for every $j \leq k-1$, and suppose by contradiction that there exists a point $x^0 \in \partial\Omega \cap \bar{W}$ such that, without loss of generality,

$$(-1)^k \frac{\partial^k}{\partial \tilde{\nu}^k}(a-b)(x^0) > 0.$$

Let m be a positive integer to be chosen later on, and let r_0 be accordingly defined as in Lemma 3.5. Let $z_\sigma = x^0 + \sigma\tilde{\nu}$, $\sigma > 0$, and $\rho > 0$ be chosen as we did in Theorem 2.2. Possibly choosing a smaller value of ρ, by (3.30) and the representation (3.36), Taylor's formula gives us

$$(a-b)(x) \geq \frac{1}{2}\frac{(-s)^k}{k\,!}\frac{\partial^k}{\partial \tilde{\nu}^k}(a-b)(x^0) \quad \text{for every } x \in U.$$

We intend to consider again the formula (3.18) with u, v being the singular solutions chosen in the proof of Theorem 2.2. For almost every $x \in \Omega$ we have

$$A(x, a(x)) - A(x, b\,(x)) \; = \int_{b(x)}^{a(x)} D_t A(x, t)\, dt, \tag{4.2}$$

and by the monotonicity assumption (2.5)

$$\begin{aligned}\Big(\int_{b\,(x)}^{a(x)} D_t A(x,\, t)\, dt\Big)\xi\cdot\xi \;&=\; \int_{b\,(x)}^{a(x)} D_t A(x,\, t)\xi\cdot\xi\, dt\\ &\geq \int_{b(x)}^{a(x)} E^{-1}|\,\xi\,|^2\, dt\\ &= (a-b)(x)E^{-1}|\,\xi\,|^2.\end{aligned}$$

In other words,

$$\int_{b(x)}^{a(x)} D_t A(x,\, t)\, dt \;=\; (a-b)(x)M(x) \qquad \text{for every } x\in U,$$

where the matrix M satisfies

$$M(x)\xi\cdot\xi \geq E^{-1}|\,\xi\,|^2 \qquad \text{for almost every } x\in U, \quad \text{for every } \xi\in\mathbb{R}^n.$$

By rephrasing the arguments leading to (3.35) and by using the induction hypothesis, which enables us to assume that the continuations of $a(x)$, $b(x)$ to $B_R(z_\sigma)\setminus\bar{\Omega}$ coincide, we obtain

$$M(x)\, Du\cdot Dv \geq C\,|x-z_\sigma|^{\,2-2(n+m)} \qquad \text{for almost every } x\in U. \tag{4.3}$$

From (3.18) we obtain

$$\begin{aligned}0 \;&=\; \int_\Omega (A(x,\, a)-A(x,\, b))\, Du\cdot Dv\\ &= \int_{\Omega\,\cap\, B_\rho(z_\sigma)} (A(x,\, a)-A(x,\, b))\, Du\cdot Dv\\ &\quad + \int_{\Omega\,\setminus\, B_\rho(z_\sigma)} (A(x,\, a)-A(x,\, b))\, Du\cdot Dv\\ &\geq \int_{\Omega\,\cap\, B_\rho(z_\sigma)} (a-b)(x)\, M(x)\, Du\cdot Dv\\ &\quad - C\int_{\Omega\,\setminus\, B_\rho(z_\sigma)} |\,a-b\,|\,|x-z_\sigma|^{\,2-2(n+m)}.\end{aligned}$$

Using (4.3), and provided we choose $m>\frac{k-1}{2}$,

$$\begin{aligned}0 \geq \frac{1}{2}\frac{(-1)^{\,k}}{k\,!}\frac{\partial^{\,k}}{\partial\tilde{\nu}^{\,k}}(a-b)(x^0)\int_{\Omega\,\cap\, B_\rho(z_\sigma)} (d(x,\,\partial\Omega))^{\,k}\,|x-z_\sigma|^{\,2-2(n+m)}\\ - C\int_{\Omega\,\setminus\, B_\rho(z_\sigma)} |\,a-b\,|\,|x-z_\sigma|^{\,2-2(n+m)}\,,\end{aligned}$$

and therefore,

$$\frac{(-1)^k}{k\,!}\frac{\partial^{\,k}}{\partial\tilde{\nu}^{\,k}}(a-b)(x^0) \leq C\,\sigma^{\,n+2m-2-k}\,.$$

Picking m such that $n+2m-2-k>0$ and letting $\sigma \to 0$, we are led to a contradiction. Then we can conclude that (4.1) holds, and consequently,

$$D^{\,j}(a-b) \;=\; 0 \qquad \text{on } \partial\Omega\cap W \quad \text{for every } j \le k. \tag{4.4}$$

Finally, (2.17) follows from (2.15) and (3.40). □

Proof of Theorem 2.4. It suffices to prove iteratively that $a\;=\;b$ on each A_j. This is obtained by the argument developed in [A, Proof of Corollary 1.1], which is based on the result of uniqueness at the boundary on the analytic continuation of $a-b$ within each A_j and on the Runge approximation theorem (see [KV2] and also [G, Theorem 2.4]). We remark that this last theorem requires only the Lipschitz continuity of the conductivity $A(x,\,a(x))$. Therefore, without need of higher order differentiability on $A(x,\,t)$, the method in [A, Proof of Corollary 1.1] also applies to the present situation. □

5. The one-eigenvalue-problem. Kohn and Vogelius have considered the case in which the $n-1$ eigenvalues and eigenvectors of a conductivity matrix A are known (see [KV1]). Their result is the following theorem.

THEOREM 5.1. *Let A, B be two symmetric, positive definite matrices with entries in $L^{\infty}(\Omega)$, and let $\{\lambda_{\,j}\}_{j\,=\,1,\dots\,,n}$, $\{\bar\lambda_{\,j}\}_{j\,=\,1,\dots\,,n}$ and $\{e_j\}_{j=1,\dots\,,n}$, $\{\bar e_j\}_{j=1,\dots\,,n}$ be the corresponding eigenvalues and eigenvectors. For $x^0\in\partial\Omega$, let B be a neighborhood of x^0 relative to $\bar\Omega$, and suppose that*

$$A,\,B\in C^{\infty}(B) \quad \textit{and } \partial\Omega\cap B \textit{ is } C^{\infty}, \tag{5.1}$$

$$e_j \;=\; \bar e_j, \quad \lambda_j \;=\; \bar\lambda_j \quad \textit{in } B, \quad 1\le j\le n-1\,, \tag{5.2}$$

$$e_n(x^0)\cdot\nu(x^0) \;\neq\; 0\,, \tag{5.3}$$

$$\Lambda_A(\phi) \;=\; \Lambda_B(\phi) \quad \textit{for every } \phi\in H^{\frac12}(\partial\Omega) \textit{ with } \operatorname{supp}(\phi)\subset B\cap\partial\Omega. \tag{5.4}$$

Then

$$D^k\lambda_n(x^0) \;=\; D^k\bar\lambda_n(x^0) \quad \textit{for any } k\;\ge 0. \tag{5.5}$$

Let us rephrase the problem of Kohn and Vogelius in terms of our setting. Letting $a(x)$ be the nth eigenvalue, the conductivity matrix A has the structure

$$A \;=\; A(x,\,a(x)), \quad \text{where}$$

$$A(x,\,t) \;=\; R(x)\begin{pmatrix} \lambda_1(x) & 0 & \dots & 0 \\ 0 & \lambda_2(x) & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \lambda_{n-1}(x) \\ 0 & 0 & \dots & t \end{pmatrix} R^{\top}(x),$$

where $\lambda_1(x), \dots, \lambda_{n-1}(x)$ are given positive functions and, for each x, $R(x)$ is a known orthogonal matrix. In this case, we have

$$D_{\,t}A(x,\,t) \;=\; R(x)\begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} R^{\top}(x).$$

We observe that the property (2.5) of monotonicity is not satisfied in this case; in fact, the following equality holds:

$$D_tA(x,\,t)\,\xi\cdot\xi \;=\; |(R^{\top}(x)\,\xi)_n|^{\,2} \tag{5.6}$$

for every $x\in\Omega$, $t\in[\lambda^{-1},\,\lambda]$, $\xi\in\mathbb{R}^n$, where the subscript n denotes the nth component.

However, it is possible to modify our previous arguments in order to prove theorems analogous to Theorems 2.1–2.4 as follows. Notice that condition (5.3) is not needed here.

CLAIM. *Theorems* 2.1, 2.2, 2.3, 2.4 *continue to hold if the monotonicity assumption* (2.5) *is replaced by*

$$E^{-1}|(R^{\top}(x)\,\xi)_n|^{\,2}\le D_{\,t}\,A(x,\,t)\,\xi\cdot\xi\le E\,|(R^{\top}(x)\,\xi)_n|^{\,2} \tag{5.7}$$

for almost every $x\in\Omega$, for every $t\in[\lambda^{-1},\,\lambda]$, for every $\xi\in\mathbb{R}^n$, where $R\,=\,R(x)$ is a given orthogonal matrix depending on the space variable x.

Proof of the Claim. For the sake of brevity, we shall point out only the crucial modification in the proof of Theorem 2.1, since the corresponding changes in the subsequent theorems follow by straightforward adaptations.

Let us recall that we have obtained

$$\begin{aligned}
&\int_{B_{\,\rho}(z_\sigma)\,\cap\,\Omega}\frac{J_b^{\,2}(A(x^0,\,a(x^0))-A(x^0,\,b(x^0)))\,J_a^{\,2}(x-z_\sigma)\cdot(x-z_\sigma)}{|J_{\,a}(x-z_\sigma)|^{\,n}\,|J_{\,b}(x-z_\sigma)|^{\,n}}\\
&\le C\int_{B_{\,\rho}(z_\sigma)\,\cap\,\Omega}|x-z_\sigma|^{\,2-2n}\,|x-x^0|^{\,\beta}\\
&\quad+\,C\int_{B_{\,\rho}(z_\sigma)\,\cap\,\Omega}|x-z_\sigma|^{\,2-2n+\,\alpha}\\
&\quad+\int_{\Omega\,\setminus\,B_\rho(z_\sigma)}|A(x,\,a)-A(x,\,b)|\,|x-z_\sigma|^{\,2-2n}\\
&\quad+\|\,\Lambda_{A(x,\,a)}-\Lambda_{A(x,\,b)}\,\|_*\|\,u\,\|_{H^{\frac12}(\partial\,\Omega)}\|\,v\,\|_{L^{\frac12}(\partial\,\Omega)}\,.
\end{aligned}$$

Notice that the monotonicity assumption was used for obtaining a lower bound on the left-hand side. Assuming (5.7), we proceed as follows:

$$\begin{aligned}
&J_b^{\,2}(A(x^0,\,a)-A(x^0,\,b))\,J_a^{\,2}(x-z_\sigma)\cdot(x-z_\sigma)\\
&\ge\Big((A(x^0,\,b))^{-1}-(A(x^0,\,a))^{-1}\Big)(x-z_\sigma)\cdot(x-z_\sigma)\\
&\quad-C\,\sigma^{\,\beta}(a(x^0)-b(x^0))\,|x-z_\sigma|^2\\
&=\left(\int_{a(x^0)}^{b(x^0)}D_{\,t}(A(x^0,\,t))^{-1}\,dt\right)(x-z_\sigma)\cdot(x-z_\sigma)\,dt\\
&\quad-C\,\sigma^{\,\beta}\,(a(x^0)-b(x^0))\,|x-z_\sigma|^2.
\end{aligned} \tag{5.8}$$

And now (assuming without loss of generality $R(x^0) = I$)

$$\begin{aligned}
&\int_{a(x^0)}^{b(x^0)} D_t A^{-1}\,\xi\cdot\xi\,dt\\
&= \int_{b(x^0)}^{a(x^0)} A^{-1}\,(D_t A)\,A^{-1}\,\xi\cdot\xi\,dt\\
&= \int_{b(x^0)}^{a(x^0)} (D_t A)\,A^{-1}\xi\cdot A^{-1}\,\xi\,dt\\
&\geq C\int_{b(x^0)}^{a(x^0)} \Big(\xi\cdot e_n\Big)^{2}\,dt\\
&= \ C\,(a(x^0)-b(x^0))\,(\xi\cdot e_n)^2,
\end{aligned}$$

where e_n denotes the nth coordinate unit vector. Then we obtain

$$\begin{aligned}
&J_b^{\,2}(A(x^0,\,a)-A(x^0,\,b))\,J_a^{\,2}\,(x-z_\sigma)\cdot(x-z_\sigma)\\
&\geq C\,(a(x^0)-b(x^0))\,|\,(x-z_\sigma)_n\,|^{\,2}\\
&-C\,\sigma^{\,\beta}\,(a(x^0)-b(x^0))\,|x-z_\sigma|^2,
\end{aligned}$$

and hence

$$\begin{aligned}
&\|\,a-b\,\|_{\,L^{\,\infty}(\partial\,\Omega)}\ \int_{B_{\,\rho}(z_\sigma)\,\cap\,\Omega}\frac{(x-z_\sigma)_n^2}{|x-z_\sigma|^{\,2n}}\\
&\leq C\Big\{\int_{B_\rho(z_\sigma)\,\cap\,\Omega}|x-z_\sigma|^{\,2-2n}\,|x-x^0|^{\,\beta}\\
&\qquad+\int_{B_\rho(z_\sigma)\,\cap\,\Omega}|x-z_\sigma|^{\,2-2n+\,\alpha}\\
&\qquad+\int_{B_\rho(z_\sigma)\,\cap\,\Omega}\sigma^{\,\beta}\,|x-z_\sigma|^{\,2-2n}\\
&\qquad+\int_{\Omega\,\setminus\,B_\rho(z_\sigma)}|A(x,\,a)-A(x,\,b)\,|\,|x-z_\sigma|^{\,2-2n}\\
&\qquad+\|\,\Lambda_{A(x,\,a)}-\Lambda_{A(x,\,b)}\,\|_*\|\,u\,\|_{H^{\frac{1}{2}}(\partial\,\Omega)}\|\,v\,\|_{H^{\frac{1}{2}}(\partial\,\Omega)}\Big\}.
\end{aligned}\tag{5.9}$$

We need to estimate the quantity $\int_{B_{\,\rho}(z_\sigma)\,\cap\,\Omega}\frac{(x-z_\sigma)_n}{|x-z_\sigma|^{\,2n}}$ from below.

Let $B_r(P)$ be a ball with center P on the axis passing through z_σ and x^0 and radius $r\leq\min\Big\{\frac{r^0}{2}-\frac{3}{4}\sigma,\,\sigma\Big\}$. In this case, we have that $B_r(P)\subset\subset(B_\rho(z_\sigma)\cap\Omega)$, and we consider a cube Q inscribed in $B_r(P)$ with an axis parallel to e_n. Then we can compute

$$\int_Q\frac{|(x-z_\sigma)_n|^{\,2}}{|x-z_\sigma|^{\,2n}}\geq C\sigma^{\,2-n},\tag{5.10}$$

and next the proof can proceed as before. □

Acknowledgment. The authors wish to express their gratitude to Professor W. R. B. Lionheart for fruitful conversations on the subject of this paper.

REFERENCES

[A] G. Alessandrini, *Singular solutions of elliptic equations and the determination of conductivity by boundary measurements*, J. Differential Equations, 84 (1990), pp. 252–272.

[G] R. Gaburro, *Sul Problema Inverso della Tomografia da Impedenza Elettrica nel Caso di Conduttivitá Anisotropa*, Tesi di Laurea in Matematica, Università degli Studi di Trieste, Trieste, Italy, 1999.

[KV1] R. Kohn and M. Vogelius, *Identification of an unknown conductivity by means of measurements at the boundary*, in Inverse Problems, SIAM-AMS Proc. 14, AMS, Providence, 1984, pp. 113–123.

[KV2] R. Kohn and M. Vogelius, *Determining conductivity by boundary measurements* II. *Interior results*, Comm. Pure Appl. Math., 38 (1985), pp. 643–667.

[L] W. R. B. Lionheart, *Conformal uniqueness results in anisotropic electrical impedance imaging*, Inverse Problems, 13 (1997), pp. 125–134.

[LaU] M. Lassas and G. Uhlmann, *On Determining a Riemannian Manifold from the Dirichlet-to-Neumann Map*, preprint, 1999.

[LeU] J. M. Lee and G. Uhlmann, *Determining anisotropic real-analytic conductivities by boundary measurements*, Comm. Pure Appl. Math., 42 (1989), pp. 1097–1112.

[M] C. B. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer, Berlin, 1966.

[N] A. Nachman, *Global uniqueness for a two dimensional inverse boundary value problem*, Ann. of Math., 142 (1995), pp. 71–96.

[S] J. Sylvester, *An anisotropic inverse boundary value problem*, Comm. Pure. Appl. Math., 43 (1990), pp. 201–32.

[SU] J. Silvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary valued problem*, Ann. of Math., 125 (1987), pp. 153–169.

SIAM J. MATH. ANAL.
Vol. 33, No. 1, pp. 172–193

DYNAMICS OF AN INTERIOR SPIKE IN THE GIERER–MEINHARDT SYSTEM*

XINFU CHEN† AND MICHAŁ KOWALCZYK‡

Abstract. We study the dynamics of an interior spike of the Gierer–Meinhardt system. Under certain assumptions on the domain size, the diffusion coefficients, and the decay rates, we prove that the velocity of the center of the spike is proportional to the negative gradient of $R(\xi,\xi)$, where $R(x,\xi)$ is the regular part of the Green's function of the Laplacian with the Neumann boundary condition. Hence, an interior spike moves towards local minima of $R(\xi,\xi)$ and therefore stays as an interior spike forever. This dynamics is fundamentally different from that of the shadow Gierer–Meinhardt system where an interior spike moves towards the closest point on the boundary.

Key words. Gierer–Meinhardt system, activator-inhibitor reaction, spikes, spike dynamics

AMS subject classifications. 35B25, 35C20, 35J60, 35K99, 92C15, 92C40

PII. S0036141099364954

1. Introduction.

1.1. The Gierer–Meinhardt system.

We consider the Gierer–Meinhardt system, for $A = A(x,t)$ and $H = H(x,t)$,

$$\begin{cases} A_t = D_A \Delta A - k_A A + l_A A^2/H, & x \in \delta\Omega = \{\delta z \mid z \in \Omega\}, t > 0, \\ H_t = D_H \Delta H - k_H H + l_H A^2, & x \in \delta\Omega, t > 0, \\ \partial_n A = 0 = \partial_n H, & x \in \partial(\delta\Omega), t > 0, \\ A(x,0) = A_0(x) > 0, \quad H(x,0) = H_0(x) > 0, & x \in \delta\Omega. \end{cases} \tag{1.1}$$

Here $\Omega \subset \mathbb{R}^N$, $N = 2, 3$, is a bounded domain with C^3 boundary and *unit volume*, Δ is the Laplace operator, ∂_n is the exterior normal derivative, and δ is the size of the physical domain. System (1.1) was proposed in [6] (see also [12]) as a model for biochemical reactions of activator-inhibitor type in which a short-range substance, the activator A, promotes its own production as well as that of a rapidly diffusing antagonist, the inhibitor H.

In this paper, we assume that D_A, D_H, k_A, and k_B, representing the diffusion coefficients and the decaying rates of species A and H, are positive constants and satisfy

$$\frac{k_A}{k_H} \ll 1, \qquad \frac{D_A}{k_A} \ll \delta^2 \ll \frac{D_H}{k_H}. \tag{1.2}$$

These conditions reflect the following scenario: (i) the half-life ($\ln 2/k_A$) of the activator A is much longer than that of the inhibitor H; (ii) with respect to the size δ

*Received by the editors December 10, 1999; accepted for publication (in revised form) November 27, 2000; published electronically April 24, 2001.
http://www.siam.org/journals/sima/33-1/36495.html

†Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, PA 15260 (xinfu@pitt.edu). The work of this author was supported by National Science Foundation grant DMS-9971043.

‡Carnegie Mellon University, Department of Mathematical Sciences, Pittsburgh, PA 15213 (kowalcyk@asdf1.math.cmu.edu). The work of this author was supported by NSF Mathematical Sciences Postdoctoral Fellowship DMS-9705972.

of the domain and the half-life of the species, the diffusion rate $D_A/(k_A\delta^2)$ of A is small whereas that of H is large; namely, regional population differences of A are not easily evened out in the life time of the component A, whereas regional population differences of H are almost instantaneously evened out by the diffusion. In such a scenario, a local increase in the concentration of the activator will be further amplified (due to the $l_A A^2/H$ term), forming regions with high concentration of the activator surrounded by the "sea" of, essentially uniformly distributed, inhibitors. We speak of *spikes* if the activator concentrates near a single point or a set of isolated points. Since stability results available at this point ([8, 17] in one dimension and [19, 20] in two dimensions) seem to suggest that in the range of parameters considered in the present paper, spikes concentrated at more than one point are unstable; therefore, here we study only the dynamics of single spike solutions.

We introduce dimensionless constants

$$\tau = \frac{k_A}{k_H}, \quad \varepsilon^2 = \frac{D_A}{k_A\delta^2}, \quad D = \frac{D_H}{k_H\delta^2}, \tag{1.3}$$

and we rescale the independent and dependent variables via

$$t \mapsto \frac{t}{k_A}, \quad x \mapsto \delta x, \quad A \mapsto \frac{k_H}{l_H\varepsilon^N}A, \quad H \mapsto \frac{k_A l_B}{k_H l_A}\varepsilon^N H. \tag{1.4}$$

Then the Gierer–Meinhardt system (1.1) takes the nondimensional form

$$\begin{cases} A_t = \varepsilon^2\Delta A - A + f(A,H), & x \in \Omega, t > 0, \\ \tau H_t = D\Delta H - H + \varepsilon^{-N}g(A), & x \in \Omega, t > 0, \\ \partial_n A = 0 = \partial_n H, & x \in \partial\Omega, t > 0, \\ A(x,0) = A_0(x) > 0, \quad H(x,0) = H_0(x) > 0, & x \in \Omega, \end{cases} \tag{1.5}$$

where

$$f(A,H) = A^2H^{-1}, \qquad g(A) = A^2. \tag{1.6}$$

Formally, as $D \to \infty$ one obtains the following *shadow* Gierer–Meinhardt system, for $A = A(x,t)$ and $H = H(t)$:

$$\begin{cases} A_t = \varepsilon^2\Delta A - A + f(A,H), & x \in \Omega, t > 0, \\ \tau H_t = -H + \varepsilon^{-N} \fint_\Omega g(A), & x \in \Omega, t > 0, \\ \partial_n A = 0 = \partial_n H, & x \in \partial\Omega, t > 0, \\ A(x,0) = A_0(x) > 0, \quad H(0) = \fint_\Omega H_0(x)dx > 0, & x \in \Omega. \end{cases} \tag{1.7}$$

Note that steady states to (1.7), after the change of variables $y = x/\varepsilon$, are solutions to

$$\Delta_y A - A + f(A, \fint_{\Omega_\varepsilon} A\,dy) = 0, \quad y \in \Omega_\varepsilon := \varepsilon^{-1}\Omega. \tag{1.8}$$

In recent years there has been much interest in studying (1.5), (1.7), and especially the associated steady state problem (1.8). In a series of papers [14, 15, 16], Ni and Takagi (also with Lin [11]) established the existence of stationary spikes (solutions to (1.8) with homogeneous Neumann boundary condition) concentrating at points of maximal mean curvature of $\partial\Omega$. We refer the readers to the recent review article by Ni [13] and the references therein for more details on this subject.

In [3] we studied the evolution of single-spike solutions to (1.7) and showed that a single, interior spike located at $\xi = \xi(t) \in \Omega$ moves toward the boundary $\partial\Omega$ with velocity

$$\dot{\xi} := \frac{d}{dt}\xi(t) \propto \nabla_\xi e^{-2\mathrm{d}(\xi)/\varepsilon}, \tag{1.9}$$

where $\mathrm{d}(\xi)$ is the distance from ξ to $\partial\Omega$. From this formula, one sees that single interior spikes for the shadow Gierer–Meinhardt system move towards their closest points on $\partial\Omega$, possibly with the exception of those which have more than one closest point on the boundary. We would like to point out that the dynamics (1.9) for the shadow Gierer–Meinhardt system (1.7) was first derived by Iron and Ward in [7], whereas in [3] we provided a rigorous proof (see also related work [19]).

In the present paper we assume that D is large, but $D \ll \varepsilon^2 e^{\max_{\xi\in\Omega}\mathrm{d}(\xi)/4\varepsilon}$. In such a case one does not expect the spike to move exponentially slowly. In fact we show that if $\varepsilon^{3-2N-\kappa} \ll D \ll \varepsilon^2 e^{\max_{\xi\in\Omega}\mathrm{d}(\xi)/4\varepsilon}$ for some $\kappa > 0$, then an interior spike moves with a velocity

$$\dot{\xi} \propto -\varepsilon^2 D^{-1}\, D_\xi R(\xi,\xi), \tag{1.10}$$

where D_ξ is the total derivative with respect to ξ, and $R(x,\xi)$ is the regular part of the Green's function for Δ with the Neumann boundary condition. Since $R(\xi,\xi) \to \infty$ as $x, \xi \to \partial\Omega$, one sees from the formula (1.10) that an interior spike moves towards local minima of $R(\xi,\xi)$ and hence stays in Ω forever.

Clearly, the dynamics (1.10) for the Gierer–Meinhardt system (1.5) is totally different from the dynamics (1.9) for the shadow Gierer–Meinhardt system (1.7).

The main purpose of this paper is to prove rigorously the asymptotic formula (1.10), following the so called *invariant manifold approach* developed by Alikakos and Fusco in [1, 2] to study motions of circular fronts (bubbles) in solutions to the Cahn–Hilliard equation.

1.2. Statement of the main result. In this paper we shall use the following notation:

$$\langle\phi\rangle := \int_\Omega \phi(x)\,dx, \qquad \langle\phi,\psi\rangle := \langle\phi\psi\rangle = \int_\Omega \phi\psi, \quad \|u\|_p := \|u\|_{L^p(\Omega)}, \quad \Omega_\varepsilon := \varepsilon^{-1}\Omega.$$

We assume that $\tau \ll 1$ and $D \gg 1$. Then we can argue from (1.5) that $H(\cdot,t)$ is almost a constant equal to $\varepsilon^{-N}\langle g(A)\rangle = \varepsilon^{-N}\langle g(A(\cdot,t))\rangle$. (Recall that the volume of Ω is 1.) Hence, it is convenient to decompose H as

$$\begin{aligned} H(x,t) &= \varepsilon^{-N}\langle g(A)\rangle + h(x,t),\\ h(x,t) &= h_0(t) + h_1(x,t),\\ h_0(t) &:= \langle H\rangle - \varepsilon^{-N}\langle g(A)\rangle,\\ h_1(x,t) &:= H(x,t) - \varepsilon^{-N}\langle g(A)\rangle - h_0(t) = H - \langle H\rangle. \end{aligned}$$

Then (1.5) can be written in terms of unknowns A, h_0, and h_1:

$$\begin{cases} A_t - \varepsilon^2\Delta A + A = f(A, \varepsilon^{-N}\langle g(A)\rangle + h_0 + h_1), & x\in\Omega, t>0,\\ \tau h_{0,t} + h_0 = -\tau\varepsilon^{-N}\langle g'(A)A_t\rangle, & t>0,\\ \tau h_{1,t} - D\Delta h_1 + h_1 = \varepsilon^{-N}[g(A) - \langle g(A)\rangle], & x\in\Omega, t>0,\\ \partial_n A = 0 = \partial_n h_1, & x\in\partial\Omega, t>0. \end{cases} \tag{1.11}$$

If we ignore h_0 and h_1 and use the stretched variable $y = x/\varepsilon$, the first equation of (1.11) becomes

$$A_t - \Delta_y A + A = f(A, \textstyle\int_{\Omega_\varepsilon} g(A)dy), \qquad y \in \Omega_\varepsilon, t > 0, \tag{1.12}$$

which is the limit, as $\tau \to 0$, of the shadow Gierer–Meinhardt system (1.7). Since $\varepsilon \ll 1$, a solution to the equation

$$-\Delta_y A + A = f(A, \textstyle\int_{\mathbb{R}^N} g(A)dy) \qquad \text{in } \mathbb{R}^N \tag{1.13}$$

will be almost a stationary solution.

With f and g given by (1.6), it is known that (1.13) has a unique, positive, radially symmetric solution, which we denote by $W(r), r = |y|$. As $W(r) \to 0$ exponentially fast as $r \to \infty$, for every $\xi \in \Omega$, $\{A = W(|x - \xi|/\varepsilon), h_0 = 0, h_1 = 0\}$ is almost an equilibrium of (1.11). In what follows, a solution with $A(x,t) \approx W(|x - \xi(t)|/\varepsilon)$, $h_0 \approx 0$ and $h_1 \approx 0$ will be called a *spike solution* located at $\xi(t)$ at time t.

We consider only spikes that initially stay away from the boundary. To this end we define $\mathrm{d}(\xi) =$ distance from ξ to $\partial\Omega$ and let μ be a parameter in the range $\max_{\xi\in\Omega} \mathrm{d}(\xi) > \mu > 4\varepsilon \log(D\varepsilon^{-2})$. We set

$$\Omega^\mu = \{\xi \in \Omega \mid \mathrm{d}(\xi) > \mu\}. \tag{1.14}$$

Observe that if D satisfies $\varepsilon^{3-2N-\kappa} \ll D \ll e^{\max_{\xi\in\Omega} \mathrm{d}(\xi)/8\varepsilon}$ for some $\kappa > 0$, then $\Omega^\mu \neq \emptyset$ for all sufficiently small ε.

It is convenient to work with approximate solutions to (1.11) which have compact support. Hence, we modify $W(|y|)$ and $W(|x - \xi|/\varepsilon)$ into compactly supported functions $W^\varepsilon(y)$ and $w^\varepsilon(x, \xi)$ as follows. Let $\zeta(s)$ be a cutoff function such that $\zeta = 1$ if $|s| < 1/2$, $\zeta = 0$ if $|s| > 1$, and $|D^n\zeta| \le 2^{n+1}$, $n = 1, 2, 3$. We define

$$\begin{aligned} &W^\varepsilon(y) = W(|y|)\zeta(|y|\varepsilon/\mu), && y \in \mathbb{R}^N, \\ &w^\varepsilon(x,\xi) = W^\varepsilon(|x-\xi|/\varepsilon) = W(|x-\xi|/\varepsilon)\zeta(|x-\xi|/\mu), && x, \xi \in \Omega. \end{aligned} \tag{1.15}$$

We define *the approximate invariant manifold* $\mathcal{M}$ by

$$\mathcal{M} = \{w^\varepsilon(\cdot, \xi) \mid \xi \in \bar{\Omega}\}. \tag{1.16}$$

It is known (see Lemma 3.1 to follow) that there exists a positive constant $c_0 > 0$ (depending only on Ω) such that if $\mathrm{dist}(A(\cdot,t), \mathcal{M}) \le c_0\varepsilon^{N/2}$, then we can uniquely decompose $A(\cdot, t)$ as

$$A(x,t) = w^\varepsilon(x, \xi(t)) + \phi(x,t), \quad \xi(t) \in \bar{\Omega}, \quad \|\phi(\cdot,t)\|_2 = \mathrm{dist}(A(\cdot,t),\mathcal{M}) := \inf_{w\in\mathcal{M}} \|A - w\|_2. \tag{1.17}$$

We define

$$T^* := \sup\{T > 0 \mid \mathrm{dist}(A(\cdot,t), \mathcal{M}) \le c_0\varepsilon^{N/2},\ \xi(t) \in \Omega^\mu \ \ \forall\, t \in [0, T]\ \}. \tag{1.18}$$

THEOREM 1.1 (quasi-invariance of the manifold $\mathcal{M}$). *Let $\kappa \in (0, 1/8)$ be any fixed constant. Let ε, τ, D, and μ be positive parameters such that there hold the relations*

$$0 < \varepsilon < 1, \quad 0 < \tau < 1, \quad \varepsilon^{2-N-2\kappa} < D, \quad 4\varepsilon \log(D\varepsilon^{-2}) < \mu < \max_{\xi\in\Omega} \mathrm{d}(\xi). \tag{1.19}$$

Let (A, h_0, h_1) *be solutions to* (1.11) *with initial values* $h_0(0), h_1(\cdot, 0)$, *and* $A(\cdot, 0) = w^\varepsilon(\cdot, \xi_0) + \phi(\cdot, 0)$, *where* $\langle h_1(\cdot, 0)\rangle = 0$, $\xi_0 \in \Omega^\mu$, *and* $\|\phi(\cdot, 0)\|_2 = \operatorname{dist}(A(\cdot, 0), \mathcal{M})$. *Assume that*

$$|h_0(0)| + \|h_1(\cdot, 0)\|_\infty + \varepsilon^{-N/2}\|\phi(\cdot, 0)\|_2 \le D^{-1}\varepsilon^{2-N-\kappa}. \tag{1.20}$$

Decompose A *as in* (1.17) *in the interval* $[0, T^*]$ *with* T^* *defined as in* (1.18).

There exist small positive constants ε_0 *and* τ_0 *and a positive constant* C_0, *all of which depend only on* Ω *and* κ, *such that if* $\varepsilon \in (0, \varepsilon_0]$ *and* $\tau \in (0, \tau_0]$, *then*

$$|h_0(t)| + \|h_1(\cdot, t)\|_\infty + \varepsilon^{-N/2}\|\phi(\cdot, t)\|_2 \le C_0 D^{-1}\varepsilon^{2-N-\kappa} < c_0 \qquad \forall t \in (0, T^*). \tag{1.21}$$

In addition, either $T^* = \infty$ *or* $\mathrm{d}(\xi(T^*)) = \mu$ *(i.e.,* $\xi(T^*) \in \partial\Omega^\mu$*).*

To describe the dynamics of the spike (and therefore show that $T^* = \infty$), we introduce the Green's function $G(x, \xi)$ of Δ with the Neumann boundary condition; i.e, for each $\xi \in \Omega$, $G(\cdot, \xi)$ solves

$$\begin{cases} -\Delta_x G(x, \xi) = \delta(x - \xi) - 1 & \text{in } \Omega, \\ \partial_n G = 0 & \text{on } \partial\Omega, \\ \int_\Omega G(x, \xi)\, dx = 0. & \end{cases} \tag{1.22}$$

Let $\Gamma(x) = -(2\pi)^{-1}\log|x|$ for $N = 2$ and $= (4\pi|x|)^{-1}$ for $N = 3$ be the fundamental solution of Δ and let $R(x, \xi) := G(x, \xi) - \Gamma(x - \xi)$ be the regular part of the Green's function.

Theorem 1.2 (dynamics of an interior spike). *In addition to the assumptions of the previous theorem we assume that*

$$\|h_1(\cdot, 0) - (D\Delta)^{-1}\{[w^\varepsilon(\cdot, \xi_0)]^2 - \langle (w^\varepsilon(\cdot, \xi_0)^2)\rangle\}\|_\infty \le D^{-2}\varepsilon^{4-2N-2\kappa}.$$

The following formula holds true:

$$\dot{\xi} = \alpha_0 D^{-1}\varepsilon^2 \left(-D_\xi R(\xi, \xi) + O(\varepsilon)\mathrm{d}(\xi)^{-N} + O(\varepsilon^{3-2N-2\kappa}D^{-1})\right) \quad \forall\, t \in (0, T^*), \tag{1.23}$$

where α_0, *given in* (2.13) *below, is a positive constant depending only on* N.

Consequently, if we further assume that $D \ge \varepsilon^{2-N-3\kappa}$ *and* μ *is sufficiently small, then* $T^* = \infty$ *and* $\xi(t) \in \Omega^\mu$ *for all* $t > 0$.

Remark 1.1. Condition (1.19) implicitly imposes an upper bound on D:

$$D < \varepsilon^2 e^{\mu/(4\varepsilon)} < \varepsilon^2 e^{\max_{\xi\in\Omega} \mathrm{d}(\xi)/(4\varepsilon)}.$$

If D is too large, say, $\log(D) > 1/\varepsilon$, then (1.5) should be considered as a small perturbation of the shadow Gierer–Meinhardt system, and therefore the dynamics (1.9) should prevail.

Intuitively, for any given $\xi \in \Omega$, making the magnitudes of the right-hand sides of (1.9) and (1.10) equal should give us the critical size of D to determine which dynamics dominates. We believe that when D is exponentially large, i.e., $\log(D) = O(1/\varepsilon)$, our analysis in [3] and the analysis presented in this paper can be combined to obtain the leading order expansion of the velocity of motion of single interior spikes, which somehow should be the sum of the right-hand sides of (1.10) and (1.9).

Remark 1.2. Our lower bound $\varepsilon^{2-N-2\kappa}$ for the magnitude of D for the quasi-invariance of the manifold $\mathcal{M}$ in Theorem 1.1 is possibly sharp. Indeed, it is proved in [20] that when $N = 2$ and $D = 1$, the stationary spike attains a maximum of the order $O(|\ln\varepsilon|)$ as $\varepsilon \to 0$, whereas in our case the spikes remain bounded by a constant independent on ε.

Remark 1.3. One of the main points of our paper is to study the dynamics of a single spike in the case when $D = D(\varepsilon)$ and τ is a small parameter *independent* on ε or D. In this context we refer the reader to [3] where the key spectral estimate (Lemma 3.2 to follow) is established. Although we believe that this estimate is true for more general systems than the one considered here, for instance, with A^2/H replaced by A^p/H^q and A^2 replaced by A^r/H^s, where $qr/[(p-1)(s+1)] > 1$; however, at the moment such results are only available in the one-dimensional case [17]. Given higher dimensional generalization of the spectral estimate in [17], one could easily adopt our method to study the dynamics of interior spikes for the more general Gierer–Meinhardt system.

Remark 1.4. One notices that taking smaller κ in our theorems makes the results stronger. Nevertheless, we cannot take $\kappa = 0$. We expect terms involving $\log\varepsilon$ will come up if we set $\kappa = 0$.

In (1.23), the term $O(\varepsilon^{3-2N-2\kappa}D^{-1})$ does not match with the combination $\mathcal{E}_\kappa := \varepsilon^{2-N-\kappa}D^{-1}$. We believe that the actual size of this term should be $O(\mathcal{E}_\kappa)$. To prove this, one needs an approximation better than approximating H by a constant function. This could, for instance, be accomplished by finding an ansatz for H from the equation $D\Delta H - H + \varepsilon^{-N}W^2 = 0$ (c.f. [20]).

Remark 1.5. We observe that since w^ε is bounded by a constant independent on ε, therefore the assumptions on $h_1(\cdot, 0)$ in Theorems 1.1 and 1.2 can be satisfied simultaneously.

Remark 1.6. When $N = 2$ and Ω is a disk of radius $1/\sqrt{\pi}$ (so area of Ω is 1), we have an explicit formula for $R(x,\xi)$. Indeed, identifying points as complex numbers, the Green's function is given by

$$G(z,\xi) = -\frac{1}{2\pi}(\ln|z-\xi| + \ln|\bar{\xi}z - 1/\pi|) + \frac{1}{4}(|z|^2 + |\xi|^2) + K_0, \quad K_0 = -\frac{3}{4\pi}\ln\pi + \frac{5}{16\pi}\ .$$

It then follows that

$$R(\xi,\xi) = -\frac{1}{2\pi}\ln\left||\xi|^2 - 1/\pi\right| + \frac{1}{2}|\xi|^2 + K_0, \qquad D_\xi R(\xi,\xi) = \frac{2-\pi|\xi|^2}{1-\pi|\xi|^2}\,\xi \quad \forall \xi \in \Omega.$$

Hence, a spike will move towards the origin in the radial direction.

For more explicit formulae of the regular part $R(x,\xi)$ of the Green's function of certain other domains, see Fraenkel [5].

Later, in Lemma 3.5, we shall show that for any smooth domain Ω, $|D_\xi R(\xi,\xi)| \propto \mathrm{d}(\xi)^{1-N}$ as $\xi \to \partial\Omega$, so that $D_\xi R(\xi,\xi)$ is the leading order term in (1.23).

In the next section we shall formally derive the dynamics of $\xi(t)$. Then in the subsequent sections, we verify the dynamics rigorously.

In what follows, we shall always assume that ε and τ are small positive constants and that D, μ satisfy (1.19).

2. Formal derivation of the dynamics.

To better explain our idea of the proof, here we first provide a formal derivation of the dynamics (1.23).

Let $W(r)$ be the solution to (1.13) and $W^\varepsilon, w^\varepsilon$ be functions defined in (1.15). We define

$$\begin{aligned} \sigma^\varepsilon &= \textstyle\int_{\mathbb{R}^N} g(W^\varepsilon(|y|))\,dy = \varepsilon^{-N}\int_{|x-\xi|<\mu} g(w^\varepsilon(x,\xi))\,dx\,, \\ r^\varepsilon = r^\varepsilon(x,\xi) &= \varepsilon^2\Delta w^\varepsilon - w^\varepsilon + f(w^\varepsilon,\sigma^\varepsilon) \\ &= \Delta_y W^\varepsilon - W^\varepsilon + f(W^\varepsilon,\sigma^\varepsilon)|_{y=(x-\xi)/\varepsilon}\,. \end{aligned} \tag{2.1}$$

Since $W(y)$ decays to zero exponentially fast as $|y| \to \infty$, we can regard r^ε as zero. (On the contrary, for the shadow Gierer–Meinhardt system, r^ε is the main term forcing a spike to move towards the boundary.)

If we decompose $A(x,t)$ as in (1.17), then the first equation of (1.11) can be written as

$$\sum_{j=1}^N w^\varepsilon_{\xi_j}\dot\xi_j + \phi_t = \mathcal{L}^\xi\phi + f_H(w^\varepsilon,\sigma^\varepsilon)(h_0+h_1) + r^\varepsilon + \mathcal{N}, \tag{2.2}$$

where

$$\begin{aligned} \mathcal{L}^\xi\phi &= \varepsilon^2\Delta\phi - \phi + f_A(w^\varepsilon,\sigma^\varepsilon)\phi + f_H(w^\varepsilon,\sigma^\varepsilon)\varepsilon^{-N}\langle g'(w^\varepsilon)\phi\rangle, \\ \mathcal{N} &= f(w^\varepsilon+\phi, \varepsilon^{-N}\langle g(w^\varepsilon+\phi)\rangle + h_0 + h_1) - f(w^\varepsilon,\sigma^\varepsilon) \\ &\quad - f_A(w^\varepsilon,\sigma^\varepsilon)\phi - f_H(w^\varepsilon,\sigma^\varepsilon)(\varepsilon^{-N}\langle g'(w^\varepsilon)\phi\rangle + h_0 + h_1). \end{aligned} \tag{2.3}$$

Multiplying (2.2) by $w^\varepsilon_{\xi_i}$ and using

$$\int_\Omega w^\varepsilon_{\xi_i}w^\varepsilon_{\xi_j} = \delta_{ij}\langle|w^\varepsilon_{\xi_i}|^2\rangle, \qquad \int_\Omega \phi_t w^\varepsilon_{\xi_i} = -\int_\Omega \phi w^\varepsilon_{\xi_i,t} = -\sum_{j=1}^N\int_\Omega \phi w^\varepsilon_{\xi_i\xi_j}\dot\xi_j$$

(since $\phi \perp T\mathcal{M}$, $\langle\phi, w_{\xi_i}\rangle = 0$ for all i), we obtain

$$\langle|w^\varepsilon_{\xi_i}|^2\rangle\dot\xi_i - \sum_{j=1}^N\langle w^\varepsilon_{\xi_i\xi_j},\phi\rangle\dot\xi_j = \int_\Omega w^\varepsilon_{\xi_i}\mathcal{L}^\xi(\phi) + \int_\Omega f_H(w^\varepsilon,\sigma^\varepsilon)(h_0+h_1)w^\varepsilon_{\xi_i} + \int_\Omega \mathcal{N}w^\varepsilon_{\xi_i}. \tag{2.4}$$

Since w_{ξ_i} is almost in the kernel of $\mathcal{L}^\xi$ and therefore it is in the kernel of its adjoint ${\mathcal{L}^\xi}^*$, we can ignore terms involving ϕ and $\mathcal{N}$ to obtain

$$\dot\xi_i\langle|w^\varepsilon_{\xi_i}|^2\rangle \approx \int_\Omega f_H(w^\varepsilon,\sigma^\varepsilon)(h_0+h_1)w^\varepsilon_{\xi_i} = \int_\Omega (h_0+h_1)[Q(w^\varepsilon,\sigma^\varepsilon)]_{\xi_i}, \tag{2.5}$$

where

$$Q(w,\sigma) = \int_0^w f_H(s,\sigma)\,ds = -\frac{w^3}{3\sigma^2} \qquad \forall\, w,\sigma \in (0,\infty). \tag{2.6}$$

Since $Q(w^\varepsilon,\sigma^\varepsilon)_{\xi_i} = -Q(w^\varepsilon,\sigma^\varepsilon)_{x_i}$ and $Q \equiv 0$ on $\partial\Omega$, we have

$$\dot\xi_i\langle|w^\varepsilon_{\xi_i}|^2\rangle \approx \int_\Omega Q(w^\varepsilon,\sigma^\varepsilon)(h_0+h_1)_{x_i} = \int_\Omega Q(w^\varepsilon,\sigma^\varepsilon)h_{1,x_i}\,. \tag{2.7}$$

Because τ is small and D is large, the third equation of (1.11) for h_1 gives

$$h_1 \approx \varepsilon^{-N}(-D\Delta)^{-1}(g(w^\varepsilon) - \langle g(w^\varepsilon)\rangle). \tag{2.8}$$

Writing the Green's function for Δ as $G(x,x') = \Gamma(x-x') + R(x,x')$ and using the fact that $\langle G(\cdot,x')\rangle = 0$, we then have

$$\begin{aligned} h_1 &\approx \varepsilon^{-N} D^{-1} \int_\Omega G(x,x')[g(w^\varepsilon) - \langle g(w^\varepsilon)\rangle](x')\,dx' = \varepsilon^{-N} D^{-1} \int_\Omega G(x,x') g(w^\varepsilon)\,dx' \\ &= \varepsilon^{-N} D^{-1} \int_\Omega \Gamma(x-x') g(w^\varepsilon)(x')\,dx' + \varepsilon^{-N} D^{-1} \int_\Omega R(x,x') g(w^\varepsilon)(x')\,dx' m \\ &=: h_{11} + h_{12}\ . \end{aligned}$$

Therefore,

$$\dot{\xi}_i \langle |w^\varepsilon_{\xi_i}|^2 \rangle \approx \int_\Omega [Q(w^\varepsilon,\sigma^\varepsilon) h_{11,x_i} + Q(w^\varepsilon,\sigma^\varepsilon) h_{12,x_i}]\ . \tag{2.9}$$

Since w^ε and h_{11} are radially symmetric about ξ and w^ε has compact support,

$$\int_\Omega Q(w^\varepsilon,\sigma^\varepsilon) h_{11,x_i} = \int_{\mathbb{R}^N} Q(w^\varepsilon,\sigma^\varepsilon) h_{11,x_i} = 0. \tag{2.10}$$

Hence

$$\begin{aligned} \dot{\xi}_i \langle |w^\varepsilon_{\xi_i}|^2 \rangle &\approx \varepsilon^{-N} D^{-1} \int_{\Omega\times\Omega} Q(w^\varepsilon,\sigma^\varepsilon)(x) R_{x_i}(x,x') g(w^\varepsilon(x',\xi))\,dx'dx \\ &= \varepsilon^{N} D^{-1} \int_\Omega \left(\int_\Omega \varepsilon^{-N} Q(w^\varepsilon(x,\xi),\sigma^\varepsilon) R_{x_i}(x,x')\,dx \right) \varepsilon^{-N} g(w^\varepsilon(x',\xi))\,dx'\ . \end{aligned}$$

Observe that as $\varepsilon \to 0$,

$$\begin{aligned} &\varepsilon^{-N} g(w^\varepsilon(x',\xi)) \to c_1 \delta(x'-\xi), && c_1 := \textstyle\int_{\mathbb{R}^N} g(W(y))\,dy, \\ &\varepsilon^{-N} Q(w^\varepsilon(x,\xi)) \to -c_2 \delta(x-\xi), && c_2 = \textstyle\int_{\mathbb{R}^N} |Q(W(y))|\,dy, \\ &\varepsilon^{2-N} \textstyle\int_\Omega |w^\varepsilon_{\xi_i}|^2 = \int_{\Omega_\varepsilon} |W^\varepsilon_{y_i}|^2 \to c_3, && c_3 := \tfrac{1}{N} \textstyle\int_{\mathbb{R}^N} |\nabla W|^2. \end{aligned} \tag{2.11}$$

We then have

$$\dot{\xi}_i \approx -\frac{c_1 c_2}{c_3} \frac{\varepsilon^2}{D} R_{x_i}(\xi,\xi) = -\frac{c_1 c_2 \varepsilon^2}{2 c_3 D} D_\xi R(\xi,\xi) =: -\alpha_0 \varepsilon^2 D^{-1} D_\xi R(\xi,\xi), \tag{2.12}$$

by using the fact that $R(\xi,x) = R(x,\xi)$ so $\nabla_x R(\xi,\xi) = \nabla_\xi R(\xi,\xi) = \frac{1}{2} D_\xi R(\xi,\xi)$.

Finally, using the definition of g, σ, and Q we have

$$\alpha_0 := \frac{c_1 c_2}{2 c_3} = \frac{N \int_{\mathbb{R}^N} W^3}{6 \int_{\mathbb{R}^N} |\nabla W|^2 \int_{\mathbb{R}^N} W^2} = \frac{N \int_{\mathbb{R}^N} (|\nabla W|^2 + W^2)}{6 \int_{\mathbb{R}^N} |\nabla W|^2}\ . \tag{2.13}$$

Here in the last equality, we have used the identity $\int_{\mathbb{R}^N} W^3 / \int_{\mathbb{R}^N} W^2 = \int_{\mathbb{R}^N} (|\nabla W|^2 + W^2)$ obtained by integrating $0 = W(-\Delta W + W - W^2 / \int_{\mathbb{R}^N} W^2)$ over $\mathbb{R}^N$.

In what follows we shall make the derivation of (2.12) rigorous.

3. Preliminaries.

3.1. The ground state $\boldsymbol{W}$. With $f(A,H) = A^2/H$, (1.13) for W reads

$$-\Delta W + W - \frac{W^2}{\int_{\mathbb{R}^N} W^2} = 0 \qquad \text{in } \mathbb{R}^N.$$

It is well known that this equation possesses a unique, nonnegative, radially symmetric solution W (refered to as the ground state) with its unique maximum attained at the origin. In addition, there exists a positive constant K such that as $r \to \infty$,

$$|D_y^\alpha W(y)| \le K e^{-|y|}, \qquad \alpha = 0, 1, 2. \tag{3.1}$$

For more details on the ground state solution W, see [4, 10, 9, 18, 21] and the references therein.

From (3.1) one sees that there exists a positive constant C, which is independent of ε and μ, such that $r^\varepsilon(x,\xi)$ defined in (2.1) satisfies

$$|r^\varepsilon(x,\xi)| + |\nabla_\xi r^\varepsilon(x,\xi)| \le C e^{-\mu/(2\varepsilon)} \le C D^{-2} \varepsilon^4 \qquad \forall x, \xi \in \Omega,$$

since $\mu \ge 4\varepsilon \log(D\varepsilon^{-2})$.

3.2. Local coordinates near $\boldsymbol{\mathcal{M}}$. For convenience, in what follows we shall often drop the superscript and write $\sigma^\varepsilon = \sigma$, $w^\varepsilon = w$, etc.

Lemma 3.1. *There exists a positive constant c_0 depending only on Ω such that if $\varepsilon \in (0,1]$ and*

$$\operatorname{dist}(u, \mathcal{M}) = \inf_{w \in \mathcal{M}} \|u - w\|_{L^2(\Omega)} < c_0 \varepsilon^{N/2}, \tag{3.2}$$

then there exists a unique $\xi \in \bar{\Omega}$ such that

$$u = w(\cdot, \xi) + \psi, \qquad \|\psi\|_2 = \operatorname{dist}(u, \mathcal{M}). \tag{3.3}$$

Consequently, if $\xi \in \Omega$, then $\psi \perp T_\xi \mathcal{M}$, the tangent space of $\mathcal{M}$ at $w(\cdot,\xi)$; that is, $\langle \psi, w_{\xi_i}(\cdot,\xi)\rangle = 0$ for all $i = 1, \dots, N$. The standard proof of this result is left to the reader.

3.3. Eigenvalue estimates. Multiplying (2.2) by ϕ and integrating over Ω yields, after using $\phi \perp T_\xi \mathcal{M}$,

$$\frac{1}{2}\frac{d}{dt}\|\phi\|_2^2 = \langle \mathcal{L}^\xi \phi, \phi\rangle + h_0 \langle f_H, \phi\rangle + \langle h_1 f_H, \phi\rangle + \langle r^\varepsilon + \mathcal{N}, \phi\rangle, \tag{3.4}$$

where the operator $\mathcal{L}^\xi$ can be written as

$$\mathcal{L}^\xi \phi = \varepsilon^2 \Delta\phi - \phi + 2\sigma^{-1} w\phi - 2\varepsilon^{-N}\sigma^{-2}\langle w, \phi\rangle w^2, \qquad w = w^\varepsilon(\cdot,\xi), \sigma = \sigma^\varepsilon. \tag{3.5}$$

Lemma 3.2. *There exists a positive constant ν which is independent of ε and μ such that for all sufficiently small positive ε,*

$$\langle \mathcal{L}^\xi \phi, \phi\rangle \le -\nu\{\varepsilon^2 \|\nabla\phi\|_2^2 + \|\phi\|_2^2\} \qquad \forall \phi \in H^1(\Omega), \quad \phi \perp T_\xi \mathcal{M}, \quad \xi \in \Omega^\mu. \tag{3.6}$$

This lemma was first established in [3, Lemma 2.4]; for completeness we include the proof in the section 8. It is worth mentioning here that this eigenvalue estimate is the key to our whole analysis.

3.4. Some L^∞ estimates for parabolic equations.

LEMMA 3.3. *There exists a positive constant $C(\Omega)$ such that for every constant $D \geq 1$, $\tau > 0$, and $T > 0$, and functions $v_0 : \Omega \to (0,\infty)$ and $F(\cdot,\cdot) : \Omega \times (0,T) \to (0,\infty)$, the solution v to*

$$\left\{\begin{array}{ll} \tau v_t - D\Delta v + v = F(x,t) & \text{in } \Omega \times (0,T), \\ \partial_n v = 0 & \text{on } \partial\Omega \times (0,T), \\ v(\cdot,0) = v_0(\cdot) & \text{in } \Omega \times \{0\} \end{array}\right. \tag{3.7}$$

satisfies

$$\min_{\Omega\times[0,T]} v \geq \frac{1}{C(\Omega)} \min\left\{ \min_{\bar\Omega} v_0,\ \min_{[0,T]} \int_\Omega F(x,t)\,dx \right\}. \tag{3.8}$$

LEMMA 3.4. *For each $p > N/2$, there exists a positive constant $C(\Omega, p)$ such that for every positive constant T, τ and η, and every function $F \in L^\infty((0,T); L^p(\Omega))$ and $v_0 \in L^\infty(\Omega)$ with $\langle v_0\rangle = \langle F(\cdot,t)\rangle = 0$ for all t, the solution v to*

$$\left\{\begin{array}{ll} \tau v_t - \Delta v + \eta v = F(x,t) & \text{in } \Omega \times (0,T), \\ \partial_n v = 0 & \text{on } \partial\Omega \times (0,T), \\ v(x,0) = v_0(x) & \text{in } \Omega \times \{0\} \end{array}\right. \tag{3.9}$$

satisfies, for every $t_0 \in [0,T]$,

$$\|v(\cdot,t_0)\|_\infty \leq \|v_0\|_\infty + C(\Omega,p) \sup_{0\leq s\leq t_0} \|F(\cdot,s)\|_p\,. \tag{3.10}$$

We leave the proofs of the above two lemmas until the last section.

3.5. The regular part of the Green's function. We assume that Ω is a bounded domain in $\mathbb{R}^N$ ($N = 2,3$) with C^3 boundary and of unit volume. We denote by $G(x,\xi)$ the Green's function for Δ in Ω with homogeneous Neumann boundary condition, i.e., the solution to (1.22). For each $\xi \in \Omega$ sufficiently close to the boundary $\partial\Omega$, the distance function $\mathrm{d}(x)$ defined as the distance from x to $\partial\Omega$ will be smooth near ξ so that there is a unique reflection point $\xi^* = \xi - 2\mathrm{d}(\xi)\nabla_\xi \mathrm{d}(\xi)$ of ξ about $\partial\Omega$.

LEMMA 3.5. *For all ξ sufficiently close to $\partial\Omega$,*

$$G(x,\xi) = \Gamma(x-\xi) + R(x,\xi), \quad R(x,\xi) = \Gamma(x-\xi^*) + J(x,\xi), \tag{3.11}$$

where $\xi^* = \xi - 2\mathrm{d}(\xi)\nabla_\xi\mathrm{d}(\xi) \in \Omega^c$ is the unique reflection point of ξ with respect to $\partial\Omega$ and the function $J(x,\xi)$ satisfies

$$|\nabla_x J(\xi,\xi)| \leq C(\Omega)\ \mathrm{d}(\xi)^{2-N}. \tag{3.12}$$

Consequently,

$$D_\xi R(\xi,\xi) = 2\nabla_x R(\xi,\xi) = 2^{2-N}(\omega_N)^{-1}\mathrm{d}(\xi)^{1-N}[-\nabla_\xi \mathrm{d}(\xi) + O(\mathrm{d}(\xi))] \quad \text{as} \ \ \xi \to \partial\Omega, \tag{3.13}$$

where ω_N is the area of the unit sphere in $\mathbb{R}^N$.

We leave the proof until the last section of the paper.

4. L^∞ estimates.

4.1. A lower bound on H.

LEMMA 4.1. *Assume that $D > 1$, $\tau > 0$, and that the initial value $H(\cdot, 0)$ of H satisfies*

$$\|H(x,0) - \sigma^\varepsilon\|_\infty \leq \frac{\sigma^\varepsilon}{4}\ . \tag{4.1}$$

There exists a constant $C = C(\Omega) > 0$ such that for all $t \in (0, T^*)$,

$$H(x,t) \geq \frac{1}{C(\Omega)}\ . \tag{4.2}$$

Proof. First of all, the constant $\sigma^\varepsilon := \int_{\mathbb{R}^N} (W^\varepsilon(y))^2 dy = \varepsilon^{-N} \int_{|x-\xi|\leq\mu} (w^\varepsilon)^2 dx$ is bounded and positive, uniformly in $\varepsilon \in (0, 1]$.

When $t \in (0, T^*)$, $A = w + \phi$ with $\|\phi\|_2 \leq c_0 \varepsilon^{N/2}$. It then follows that

$$\bar\sigma(t) := \varepsilon^{-N} \int_\Omega g(A) = \varepsilon^{-N} \int_\Omega A^2 = \varepsilon^{-N} \int_\Omega (w+\phi)^2 \in (\sigma^\varepsilon/2, 3\sigma^\varepsilon/2) \tag{4.3}$$

(taking smaller c_0 if necessary). The assertion of the lemma then follows directly from Lemma 3.3 with $F(x,t) = \varepsilon^{-N} g(A)$. ☐

4.2. An upper bound for A.

LEMMA 4.2. *There exists a positive constant $C(\Omega)$ such that*

$$\|A\|_{\infty,\Omega\times[0,T^*]} \leq C(\Omega), \qquad \|\phi\|_{\infty,\Omega\times[0,T^*]} \leq C(\Omega)\ . \tag{4.4}$$

Proof. Set $y = x/\varepsilon$. Then (1.5) can be written as

$$A_t - \Delta_y A + A = f(A, H) \qquad \text{in } \Omega_\varepsilon \times (0, T^*).$$

Fix $p \in (N/2, 2)$. Then the local and boundary parabolic estimates yield

$$\|A\|_{\infty,\Omega\times[0,T^*]} \leq \|A(\cdot,0)\|_\infty + C \sup_{0<t\leq T^*} \sup_{y\in\Omega_\varepsilon} (\|A(\cdot,t)\|_{L^2(B_1(y)\cap\Omega_\varepsilon)} + \|f(A,H)\|_{L^p(B_1(y)\cap\Omega_\varepsilon)}), \tag{4.5}$$

where C depends only on the $C^{2+\alpha}$ norm of $\partial\Omega_\varepsilon$ and hence is bounded independently on $\varepsilon \in (0, 1]$. As $f(A, H) = A^2/H$,

$$\begin{aligned}\|f(A,H)\|_{L^p(B_1(y)\cap\Omega_\varepsilon)} &\leq \frac{\|A^2\|_{p,\Omega_\varepsilon}}{\min H} \leq C\ \|A\|_\infty^{2-2/p} \|A\|_{2,\Omega_\varepsilon}^{2/p} \\ &\leq \delta\|A\|_\infty + C(\delta)\|A\|_{2,\Omega_\varepsilon}^{2/(2-p)} = \delta\|A\|_\infty + C(\delta)\left(\int_\Omega \varepsilon^{-N} A^2 dx\right)^{1/(2-p)}.\end{aligned}$$

Taking small δ, we then obtain from (4.5) that

$$\begin{aligned}\|A\|_{\infty,\Omega\times[0,T^*]} &\leq \Big\{\|A(\cdot,0)\|_\infty + C \sup_{(0,T^*)} [\bar\sigma(t) + C(\delta)\bar\sigma^{1/(2-p)}(t)]\Big\}(1 - C\delta)^{-1} \\ &\leq C(\Omega),\end{aligned}$$

where $\bar\sigma(t)$ is as in (4.3). The proof of the lemma is complete. ☐

5. The flow in the normal space to $\mathcal{M}$.

5.1. Estimates for ϕ. With f and g given by (1.6),

$$H = \varepsilon^{-N}\langle (w+\phi)^2\rangle + h = \sigma + 2\varepsilon^{-N}\langle \phi w\rangle + \varepsilon^{-N}\|\phi\|_2^2 + h,$$

and (2.2) reads

$$\nabla_\xi w \cdot \dot{\xi} + \phi_t = r^\varepsilon + \mathcal{L}^\xi \phi - \frac{w^2}{\sigma^2} h - \mathcal{N}[w,\phi], \tag{5.1}$$

where $r^\varepsilon = \varepsilon^2 \Delta w - w + w^2/\sigma$ and

$$\mathcal{N}[w,\phi] = \frac{\phi^2}{H} - \frac{\varepsilon^{-N}\|\phi\|_2^2 w^2}{\sigma^2} - \frac{(h + 2\varepsilon^{-N}\langle w\phi\rangle + \varepsilon^{-N}\|\phi\|_2^2)w\phi}{\sigma H} + \frac{w^2(H-\sigma)^2}{\sigma^2 H}\ . \tag{5.2}$$

LEMMA 5.1. *The following estimates hold for all $t \in [0, T^*]$:*

$$|\mathcal{N}[w,\phi]| \le C\ (\ \phi^2 + \varepsilon^{-N}\|\phi\|_2^2 w^2 + h^2 w^2\ ), \tag{5.3}$$

$$\frac{1}{2}\frac{d}{dt}\|\phi\|_2^2 \le (\varepsilon^2\|\nabla\phi\|_2^2 + \|\phi\|_2^2)\left(\varepsilon^{-N/2}\|\phi\|_2 - \frac{\nu}{2}\right) + C(\nu)(\varepsilon^N\|h\|_\infty^2 + |r^\varepsilon|_\infty^2), \tag{5.4}$$

where $\nu > 0$ is the constant in Lemma 3.2.

Proof. The estimate (5.3) follows from Lemma 4.1, the bounds

$$\|\phi\|_2 \le c_0 \varepsilon^{N/2}, \qquad \|w\|_p^p = \varepsilon^N \int_{\mathbb{R}^N} (W^\varepsilon)^p dy = O(\varepsilon^N) \quad \forall\, p > 1, \tag{5.5}$$

and a straightforward calculation.

To prove (5.4), we multiply (5.1) by ϕ, integrate over Ω, and use $\phi \perp \mathcal{M}$, obtaining

$$\frac{1}{2}\frac{d}{dt}\|\phi\|_2^2 \le \langle \mathcal{L}^\xi\phi, \phi\rangle + |\langle r^\varepsilon, \phi\rangle| + |\langle w^2 h\sigma^{-2}, \phi\rangle| + |\langle \mathcal{N}[w,\phi], \phi\rangle|. \tag{5.6}$$

Let ν be the constant in Lemma 3.2. Using the bounds in (5.5), we can estimate

$$\begin{aligned}
|\langle r^\varepsilon, \phi\rangle| &\le \frac{\nu}{8}\|\phi\|_2^2 + C\|r^\varepsilon\|_\infty^2,\\
|\langle w^2 h\sigma^{-2}, \phi\rangle| &\le C\varepsilon^{N/2}\|\phi\|_2\|h\|_\infty \le \frac{\nu}{8}\|\phi\|_2^2 + C\varepsilon^N\|h\|_\infty^2,\\
|\langle \phi^2, \phi\rangle| &\le \|\phi\|_3^3 \le C\varepsilon^{-N/2}\|\phi\|_2(\varepsilon^2\|\nabla\phi\|_2^2 + \|\phi\|_2^2),\\
\varepsilon^{-N}\|\phi\|_2^2|\langle w^2\phi\rangle| &\le C\varepsilon^{-N/2}\|\phi\|_2^3,\\
|\langle h^2 w^2 \phi\rangle| &\le C\varepsilon^{N/2}\|h\|_\infty^2\|\phi\|_2 \le C\varepsilon^N\|h\|_\infty^2.
\end{aligned}$$

Substituting these estimates into (5.6) and (5.3), and using Lemma 3.2, we then obtain (5.4). □

5.2. Estimates for h. We first estimate h_1. It is convenient to further decompose $h_1 = h_{11} + h_{12}$, where

$$\begin{aligned}
&\tau D^{-1} h_{11,t} - \Delta h_{11} + D^{-1} h_{11} = \varepsilon^{-N} D^{-1}(w^2 - \langle w^2\rangle)\ ,\\
&\tau D^{-1} h_{12,t} - \Delta h_{12} + D^{-1} h_{12} = \varepsilon^{-N} D^{-1}[(2\phi w + \phi^2) - \langle 2\phi w + \phi^2\rangle].
\end{aligned}$$

Both h_{11} and h_{12} satisfy the homogeneous Neumann boundary condition, $h_{12}(\cdot, 0) \equiv 0$, and $h_{11}(x,0) = h_1(x,0) = H(x,0) - \langle H(\cdot,0)\rangle$.

LEMMA 5.2. *Assume that $h_{12}(\cdot,0) \equiv 0$ and for some $\kappa \in (0,1/4)$, $\|h_{11}(\cdot,0)\|_\infty \leq D^{-1}\varepsilon^{2-N-\kappa}$. Then there exists $C = C(\kappa,\Omega)$ such that*

$$\|h_{11}\|_{\infty,\Omega\times[0,T^*]} \leq CD^{-1}\varepsilon^{2-N-\kappa} \ , \tag{5.7}$$

$$\|h_{12}\|_{\infty,\Omega\times[0,T^*]} \leq CD^{-1}\varepsilon^{2-N-\kappa}(\varepsilon^{-N/2}\|\phi\|_2). \tag{5.8}$$

Proof. We set $p = N/(2-\kappa) \in (N/2, 2)$. Using Lemma 3.4 we obtain

$$\begin{aligned} \|h_{11}\|_{\infty,\Omega\times[0,T^*]} &\leq \|h_{11}(\cdot,0)\|_\infty + CD^{-1}\sup\nolimits_{0<s<T^*}\|\varepsilon^{-N}w^2\|_p \ , \\ \|h_{12}\|_{\infty,\Omega\times[0,T^*]} &\leq CD^{-1}\sup\nolimits_{0<s<T^*}\|\varepsilon^{-N}(2w\phi+\phi^2)\|_p \ . \end{aligned} \tag{5.9}$$

Since $\varepsilon^{-N}\|w^2\|_p \leq C\varepsilon^{-N+N/p} = C\varepsilon^{2-N-\kappa}$, the estimate (5.7) follows from (5.9).

Also, we have

$$\|\varepsilon^{-N}w\phi\|_p \leq \varepsilon^{-N}\|\phi\|_2\|w\|_{2p/(2-p)} \leq C\varepsilon^{-N+N(2-p)/(2p)}\|\phi\|_2 = C\varepsilon^{-N-\kappa}(\varepsilon^{-N/2}\|\phi\|_2) \ ,$$
$$\|\varepsilon^{-N}\phi^2\|_p \leq \varepsilon^{-N}\|\phi\|_2^{2/p}\|\phi\|_\infty^{2-2/p} \leq C\varepsilon^{2-N-\kappa}(\varepsilon^{-N/2}\|\phi\|_2)$$

since $\|\phi\|_\infty \leq C$ and $\varepsilon^{-N/2}\|\phi\|_2 \leq c_0$. The inequality (5.8) then follows from (5.9) and the preceding estimates. □

In what follows we denote

$$\mathcal{E}_\kappa = \mathcal{E}_\kappa(\varepsilon, D) = D^{-1}\varepsilon^{2-N-\kappa} \quad \text{for } \kappa \in (0,1/4) \ . \tag{5.10}$$

We shall now estimate h_0, which solves

$$h_{0,t} + \tau^{-1}h_0 = -2\varepsilon^{-N}\int_\Omega AA_t \ . \tag{5.11}$$

LEMMA 5.3. *The following estimate holds true for all $t \in [0,T^*]$,*

$$\frac{1}{2}\frac{d}{dt}h_0^2+\tau^{-1}h_0^2 \leq C(\kappa,\Omega)[\|r^\varepsilon\|_\infty^2+\|h\|_\infty^2+\varepsilon^{-N}\|\phi\|_2^2+\varepsilon^{-N}|h_0|(\varepsilon^2\|\nabla\phi\|_2^2+\|\phi\|_2^2)]. \tag{5.12}$$

Proof. Substituting

$$\int_\Omega AA_t = \int_\Omega (w+\phi)\{r^\varepsilon + \mathcal{L}^\xi\phi - \sigma^{-2}w^2h + \mathcal{N}[w,\phi]\}$$

into (5.11) and using a straightforward calculation similar to that in the proofs of Lemmas 5.1 and 5.2, we obtain (5.12). We omit the details. □

5.3. Proof of Theorem 1.1. Adding estimates (5.4) and (5.12) we obtain

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}(h_0^2 + \varepsilon^{-N}\|\phi\|_2^2) \leq (C|h_0| + C\varepsilon^{-N}\|\phi\|_2^2) - \nu/2\varepsilon^{-N}(\varepsilon^2\|\nabla\phi\|_2^2 + \|\phi\|_2^2) \\ +(C-\tau^{-1})h_0^2 + C\|h\|_\infty^2 + C\varepsilon^{-N}\|r^\varepsilon\|_\infty^2 \ . \end{aligned} \tag{5.13}$$

By Lemma 5.2,

$$\begin{aligned} \|h\|_\infty^2 = \|h_0 + h_{11} + h_{12}\|_\infty^2 &\leq 2h_0^2 + 2\|h_{11}\|_\infty^2 + 2\|h_{12}\|_\infty^2 \\ &\leq 2h_0^2 + C(\mathcal{E}_\kappa)^2(1+\varepsilon^{-N}\|\phi\|_2^2) \ . \end{aligned}$$

Since $\|r^\varepsilon\|_\infty \le Ce^{-\mu/(2\varepsilon)} < CD^{-2}\varepsilon^4$, taking τ_0, ε_0, c_0 (in Lemma 3.1) sufficiently small we obtain from (5.13) that if $\varepsilon \in (0, \varepsilon_0], \tau \in (0, \tau_0]$, then

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}(h_0^2 + \varepsilon^{-N}\|\phi\|_2^2) &\le -\frac{1}{2\tau}h_0^2 - (\nu/4 - C|h_0|)\varepsilon^{-N}(\varepsilon^2\|\nabla\phi\|^2 + \|\phi\|_2^2) + C_2(\mathcal{E}_\kappa)^2 \\ &\le -\frac{1}{C_1}(h_0^2 + \varepsilon^{-N}\|\phi\|^2) + C_2(\mathcal{E}_\kappa)^2 \qquad (5.14)\end{aligned}$$

for all $t \in [0, \hat{T}]$, where $[0, \hat{T}]$ is the maximal interval in $[0, T^*]$ in which $|h_0| \le \nu/(8C)$. Here C_1 and C_2 are constants depending only on N, κ, and Ω.

Applying Gronwall's inequality to (5.14), we conclude that there exists a constant $C = C(\Omega, \kappa, N)$ such that when $\varepsilon \in (0, \varepsilon_0]$ and $\tau \in (0, \tau_0]$,

$$h_0^2 + \varepsilon^{-N}\|\phi\|_2^2 \le C(\mathcal{E}_\kappa)^2 = C(D^{-1}\varepsilon^{2-N-\kappa})^2$$

for all $t \in (0, \hat{T}]$. Since we assume that $D > \varepsilon^{2-N-2\kappa}$, we see from the above estimate that, taking ε_0 smaller if necessary, $|h_0| \le \nu/(16C)$ and $\varepsilon^{-N/2}\|\phi\|_2 \le c_0/2$. Thus, we must have $\hat{T} = T^*$ and, by the definition of T^*, either $T^* = \infty$ or $\xi(T^*) \in \partial\Omega^\mu$. This completes the proof of Theorem 1.1. □

6. The flow in the tangent space of $\mathcal{M}$.

In what follows, we assume that $\varepsilon \in (0, \varepsilon_0]$, $\tau \in (0, \tau_0]$, and that D and μ satisfy (1.19). Then the assertion of Theorem 1.1 holds true.

6.1. The velocity.

Lemma 6.1. *For all* $t \in [0, T^*)$,

$$\dot{\xi} = \frac{\varepsilon^2}{3\sigma^2 c_3}\Big(\mathrm{I} + O(\mathcal{E}_\kappa)\Big)\Big(-\varepsilon^{-N}\langle\nabla_x h_{11}, w^3\rangle + O(\varepsilon^{-1}{\mathcal{E}_\kappa}^2)\Big), \qquad (6.1)$$

where I *is the identity matrix and* $c_3 = N^{-1}\int_{\mathbb{R}^N}|\nabla W(y)|^2 dy$.

Proof. Multiplying (5.1) by w_{ξ_j} and integrating the resulting equation over Ω yields

$$\sum_{i=1}^N \dot{\xi}_i\Big(\langle w_{\xi_i}, w_{\xi_j}\rangle - \langle w_{\xi_i\xi_j}, \phi\rangle\Big) = \langle r^\varepsilon, w_{\xi_j}\rangle + \langle\mathcal{L}^\xi\phi, w_{\xi_j}\rangle - \langle\sigma^{-2}w^2 h, w_{\xi_j}\rangle + \langle\mathcal{N}, w_{\xi_j}\rangle. \qquad (6.2)$$

Note that

$$\begin{aligned}&\langle w_{\xi_i}, w_{\xi_j}\rangle = \varepsilon^{N-2}\int_{\mathbb{R}^N} W^\varepsilon_{y_i}W^\varepsilon_{y_j}\,dy = \varepsilon^{N-2}c_3\delta_{ij}(1 + O(e^{-\mu/\varepsilon}))\ , \\ &|\langle w_{\xi_i\xi_j}, \phi\rangle| \le C\varepsilon^{N-2}(\varepsilon^{-N/2}\|\phi\|_2) \le C\varepsilon^{N-2}\mathcal{E}_\kappa\ .\end{aligned} \qquad (6.3)$$

Hence, (6.2) can be written as

$$c_3\varepsilon^{N-2}(\mathrm{I} + O(\mathcal{E}_\kappa))\dot{\xi} = \langle r^\varepsilon, w_\xi\rangle + \langle\mathcal{L}^\xi\phi, w_\xi\rangle - \langle\sigma^{-2}w^2 h, w_\xi\rangle + \langle\mathcal{N}, w_\xi\rangle. \qquad (6.4)$$

We shall now estimate each term on the right-hand side.

First of all,

$$|\langle r^\varepsilon, w_\xi\rangle| \le \|r^\varepsilon\|_\infty|\langle|w_\xi|\rangle \le Ce^{-\mu/(2\varepsilon)}\varepsilon^{N-1} \le C\varepsilon^{N-1}{\mathcal{E}_\kappa}^2. \qquad (6.5)$$

Denote

$$\begin{aligned}(\mathcal{L}^\xi)^*\psi &= \varepsilon^2\Delta\psi - \psi + 2\sigma^{-1}w\psi - 2\varepsilon^{-N}\sigma^{-2}\langle w^2\psi\rangle w\\ &= \mathcal{L}^\xi\psi - 2\sigma^{-2}\varepsilon^{-N}(\langle w^2\psi\rangle w - \langle w\psi\rangle w^2).\end{aligned}$$

Then $(\mathcal{L}^\xi)^* w_\xi = \mathcal{L}^\xi w_\xi = r^\varepsilon_\xi$. It follows that

$$|\langle\mathcal{L}^\xi\phi, w_\xi\rangle| = |\langle\phi, (\mathcal{L}^\xi)^* w_\xi\rangle| = |\langle\phi, r^\varepsilon_\xi\rangle| \le Ce^{-\mu/(2\varepsilon)}\varepsilon^{N/2}\mathcal{E}_\kappa \le \varepsilon^{N-1}{\mathcal{E}_\kappa}^2.$$

Next, from (5.3) we obtain

$$\begin{aligned}|\langle\mathcal{N}, w_\xi\rangle| &\le C\int_\Omega(\phi^2 + \varepsilon^{-N}\|\phi\|_2^2 w^2 + h^2w^2)|w_\xi|\\ &\le C\varepsilon^{N-1}[\varepsilon^{-N}\|\phi\|_2^2 + \|h\|_\infty^2] \le C\varepsilon^{N-1}{\mathcal{E}_\kappa}^2.\end{aligned}$$

As $h = h_0(t) + h_1(x,t) = h_0 + h_{11} + h_{12}$,

$$\langle\sigma^{-2}w^2h, w_\xi\rangle = \langle\sigma^{-2}w^2h_1, w_\xi\rangle = \langle\sigma^{-2}w^2h_{11}, w_\xi\rangle + \langle\sigma^{-2}w^2h_{12}, w_\xi\rangle$$

and

$$|\langle\sigma^{-2}w^2h_{12}, w_\xi\rangle| \le C\varepsilon^{N-1}\|h_{12}\|_\infty \le C\varepsilon^{N-1}{\mathcal{E}_\kappa}^2 \tag{6.6}$$

by Lemma 5.2.

Finally, observe that

$$\langle w^2h_{11}, w_\xi\rangle = -\frac{1}{3}\langle h_{11}, (w^3)_x\rangle = \frac{1}{3}\langle\nabla_x h_{11}, w^3\rangle.$$

Substituting all these estimates into (6.4) we then obtain (6.1) and complete the proof of the lemma. □

Lemma 6.2. *Under the assumptions of Theorem* 1.2, *formula* (1.23) *holds with* α_0 *given by* (2.13).

Proof. From (6.1), it suffices to estimate the term $\langle\nabla_x h_{11}, w^3\rangle$.

We write $h_{11} = h_{110} + h_{111}$, where

$$\begin{aligned}&-\Delta h_{110} = D^{-1}\varepsilon^{-N}[w^2 - \langle w^2\rangle],\\ &\tau D^{-1}h_{111,t} - \Delta h_{111} + D^{-1}h_{111} = -D^{-1}(\tau h_{110,t} + h_{110})\end{aligned} \tag{6.7}$$

with the homogeneous Neumann boundary condition for both h_{110} and h_{111}. Note that

$$h_{111}(\cdot,0) = h_{11}(\cdot,0) - h_{110}(\cdot,0) = h_1(\cdot,0) - (D\Delta)^{-1}(w(\cdot,\xi_0)^2 - \langle w^2(\cdot,\xi_0)\rangle).$$

Using the Green's function for the Δ, we have

$$\begin{aligned}h_{110}(x) &= D^{-1}\varepsilon^{-N}\int_\Omega\Gamma(x-\eta)w^2(\eta,\xi)\,d\eta + D^{-1}\varepsilon^{-N}\int_\Omega R(x,\eta)w^2(\eta,\xi)\,d\eta\\ &\overset{\text{def}}{=} \psi_1(|x-\xi|) + \psi_2(x,\xi).\end{aligned}$$

Note that $\langle w^3, \nabla_x\psi_1\rangle = 0$ so that $\langle\nabla_x h_{110}, w^3\rangle = \langle\nabla_x\psi_2, w^3\rangle$.

As $|\nabla_x \nabla_\xi R(x,\xi)| \le C(\Omega)\mathrm{d}(\xi)^{-N}$ for all $x \in \Omega$,

$$\begin{aligned}
&\left|\int_\Omega R_{x_i}(x,\eta) w^2(\eta,\xi)\,d\eta - R_{x_i}(x,\xi)\int_\Omega w^2(\eta,\xi)\,d\eta\right| \\
&\le \int_{|\eta-\xi|\le\mu} |R_{x_i}(x,\eta) - R_{x_i}(x,\xi)| w^2(\eta,\xi)\,d\eta \\
&\le \int_{|\eta-\xi|<\mu/2} |\nabla_x\nabla_\xi R|\,|\eta-\xi|\;w^2 d\eta \\
&+ \int_{\mu/2<|\eta-\xi|\le\mu} |R_{x_i}(x,\eta) - R_{x_i}(x,\xi)| w^2(\eta,\xi)\,d\eta \le C\varepsilon^{N+1}\mathrm{d}(\xi)^{-N}.
\end{aligned}$$

Similarly, since $|\nabla_x\nabla_x R(x,\xi)| \le C\mathrm{d}(\xi)^{-N}$,

$$\left|\int_\Omega R_{x_i}(x,\xi) w^3(x,\xi)\,dx - R_{x_i}(\xi,\xi)\int_\Omega w^3(x,\xi)\,d\eta\right| \le C\varepsilon^{N+1}\mathrm{d}(\xi)^{-N}.$$

It then follows that

$$\begin{aligned}
\langle \varepsilon^{-N} w^3, h_{110,\xi}\rangle &= D^{-1}\int_\Omega \varepsilon^{-N} w^3(x,\xi)\int_\Omega \varepsilon^{-N} R_{x_i}(x,\eta) w^2(\eta,\xi)\,d\eta dx \\
&= D^{-1} R_{x_i}(\xi,\xi)\int_{\mathbb{R}^N} W^2 \int_{\mathbb{R}^N} W^3 + O(\varepsilon) D^{-1}\mathrm{d}(\xi)^{-N}\,. \qquad (6.8)
\end{aligned}$$

On the other hand, by Lemma 3.4 we have

$$\begin{aligned}
\|h_{111}\|_\infty &\le \|h_{110}(\cdot,0)\|_\infty + \sup_{0<t<T^*} D^{-1}(\tau\|h_{110,t}\|_p + \|h_{110}\|_p) \\
&\le C{\mathcal{E}_\kappa}^2 + D^{-2}\sup_{0<t<T^*}(\tau\|\Delta^{-1}[\varepsilon^{-N} w\nabla_\xi w\cdot\dot{\xi}]\|_p + \|\Delta^{-1}\varepsilon^{-N} w^2\|_p) \\
&\le C{\mathcal{E}_\kappa}^2 + CD^{-2}\sup_{0<t<T^*}[(\tau|\dot{\xi}|+1)\varepsilon^{-N}\|w^2\|_p] \\
&\le C{\mathcal{E}_\kappa}^2
\end{aligned}$$

since from (5.7) and (6.1) $|\dot{\xi}| = o(1)$ as long as $t < T^*$. Thus,

$$\varepsilon^{-N}|\langle w^2 h_{111}, w_{\xi_i}\rangle| \le C\varepsilon^{-1}\|h_{111}\|_\infty \le C\varepsilon^{-1}{\mathcal{E}_\kappa}^2. \qquad (6.9)$$

Hence, from (1.23), the preceding estimates, and the definition of α_0, we obtain

$$\dot{\xi} = \alpha_0\varepsilon^2 D^{-1}\Big(\mathrm{I} + O(\mathcal{E}_\kappa)\Big)\Big(-D_\xi R(\xi,\xi) + O(\varepsilon)\mathrm{d}(\xi)^{-N} + O(\varepsilon^{3-2N-2\kappa}D^{-1})\Big).$$

Finally, notice that $|\mathcal{E}_\kappa D_\xi R(\xi,\xi)| \le C\mathrm{d}(\xi)^{1-N}\varepsilon^{2-N-\kappa}D^{-1} = C\varepsilon^{3-2N-\kappa}D^{-1}[\varepsilon/\mathrm{d}(\xi)]^{N-1}$ $\le C\varepsilon^{3-2N-2\kappa}D^{-1}$. Equation (1.23) thus follows. □

6.2. Proof of Theorem 1.2. It remains to show that $T^* = \infty$ when $D > \varepsilon^{3-2N-3\kappa}$ and μ is small.

When $\xi(t)$ is near the boundary $\partial\Omega$, we have, from (1.23) and (3.13),

$$\dot{\xi} = 2^{2-N}\alpha_0\varepsilon^2(\omega_N D)^{-1}\mathrm{d}(\xi)^{1-N}\Big(\nabla_\xi\mathrm{d}(\xi) + O(\varepsilon/\mathrm{d}(\xi)) + O(\varepsilon^{3-2N-2\kappa}D^{-1}\mathrm{d}(\xi)^{N-1})\Big).$$

It then follows that $\frac{d}{dt}\mathrm{d}(\xi(t)) > 0$ whenever $\xi(t)$ is close enough to the boundary. Consequently, $\mathrm{d}(\xi(t)) > \mu$ for all $t \in [0, T^*]$ if we take μ small enough. Therefore, by Theorem 1.1, $T^* = \infty$. □

7. Proofs of auxiliary lemmas.

Proof of Lemma 3.3. Integrating the differential equation over Ω yields

$$\tau\frac{d}{dt}\int_\Omega v+\int_\Omega v=\int_\Omega F(x,t)\,dx\geq \min_{[0,T]}\int_\Omega F(x,t)\,dx\,. \tag{7.1}$$

Gronwall's inequality then gives

$$\int_\Omega v(x,t)\,dx\geq \min\left\{\int_\Omega v_0(x)\,dx,\ \min_{[0,T]}\int_\Omega f(x,t)\,dx\right\}\,. \tag{7.2}$$

To prove (3.10), we consider two cases: (i) $t_0\in[0,\tau]$; (ii) $t_0\in[\tau,T]$.

Case (i): $t_0\in[0,\tau]$. Comparing v with a subsolution $\underline{v}=e^{-t/\tau}\min_{\bar\Omega}v_0$ gives

$$v(x,t_0)\geq \underline{v}\geq \frac{1}{e}\min_{\bar\Omega}v_0\quad \forall\, x\in\Omega. \tag{7.3}$$

Case (ii): $t_0\in[\tau,T]$. We define

$$s=D(t-t_0)/\tau+1,\qquad \tilde v=e^{(t-t_0)/\tau}v=e^{(s-1)/D}v\,.$$

Then $\tilde v_s-\Delta\tilde v\geq 0$ so that there exists $\tilde C(\Omega)$ such that

$$\min_\Omega \tilde v(\cdot,1)\geq \tilde C(\Omega)\int_\Omega \tilde v(\cdot,0)=\tilde C(\Omega)e^{-1/D}\int_\Omega v(\cdot,t_0-\tau/D); \tag{7.4}$$

namely,

$$\min_\Omega v(\cdot,t_0)\geq \tilde C(\Omega)e^{-1/D}\int_\Omega v(\cdot,t_0-\tau/D)\quad \forall t_0\in[\tau,T]. \tag{7.5}$$

Combining (7.2)–(7.5) then yields the assertion (3.10) of the lemma. □

Proof of Lemma 3.4. Since the equation is linear we can assume that $v_0\equiv 0$. In addition, by the change of variables $t'=t/\tau$ we can also assume that $\tau=1$. Furthermore, we can assume that $\eta=0$ because, by defining $\tilde v=e^{\eta t}v$ and $\tilde F=e^{\eta t}F$, if (3.10) holds for $(\tilde v,\tilde F)$, then it automatically holds for (v,F) by the assumption that $\eta\geq 0$. Thus it suffices to establish (3.10) for the case when $\tau=1$, $\eta=0$, and $v_0\equiv 0$.

Integrating the differential equation over Ω yields $\langle v\rangle=0$ for all $t\in[0,T]$.

Multiplying the differential equation by v and integrating the resulting equation over Ω gives

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_\Omega v^2+\int_\Omega|\nabla v|^2&=\int_\Omega Fv\\ &\leq \|F\|_{(2^*)'}\|v\|_{2^*}\leq C(\delta)\|F\|^2_{(2^*)'}+\delta\|v\|^2_{2^*},\end{aligned}$$

where $2^*=2N/(2-N)$ for $N\geq 3$ and any large number when $N=2$. First applying the Poincare inequality $\|v\|_{2^*}\leq C(\Omega)\|\nabla v\|_2$, then choosing $\delta=1/(2C(\Omega))$, and finally applying the Gronwall's inequality, we then obtain

$$\|v(\cdot,t)\|_2\leq C\sup_{0<s<t}\|F(\cdot,s)\|_p$$

for any $p>(2^*)'=2N/(2+N)$ when $N\geq 3$ and $p>1$ when $N=2$. The assertion of the lemma then follows from the parabolic estimate

$$\|v(\cdot,t)\|_\infty\leq C(\Omega,p)\sup_{0<s<t}(\|v(\cdot,s)\|_2+\|f(\cdot,s)\|_p)\quad \forall p\in(N/2,\infty),$$

since $v(\cdot,0)=0$. The proof of the lemma is complete. $\square$

Proof of Lemma 3.5. Since (3.13) follows directly from (3.12) and the definition of Γ and ξ^*, we need to show only (3.12).

Note that $J(x,\xi) := G(x,\xi) - \Gamma(x-\xi) - \Gamma(x-\xi^*)$ satisfies $\Delta_x J(x,\xi) = 1$ in Ω, $\partial_n J(x,\xi) = b(x,\xi)$ on $\partial\Omega$, and $\int_\Omega J(x,\xi)dx = c(\xi)$, where

$$b(x,\xi) = -\partial_n\Gamma(x-\xi) - \partial_n\Gamma(x-\xi^*), \quad c(\xi) = -\int_\Omega (\Gamma(x-\xi)+\Gamma(x-\xi^*))\,dx\ .$$

A geometric argument shows that for all $x\in\partial\Omega$ and $\xi\in\Omega$,

$$|b(x,\xi)| = \frac{1}{\omega_N}\left|\frac{(x-\xi)\cdot n(x)}{|x-\xi|^N} + \frac{(x-\xi^*)\cdot n(x)}{|x-\xi^*|^N}\right| \le C(\Omega)|\xi-x|^{2-N}. \tag{7.6}$$

Using the Green's formula and noting that $\langle G(\cdot,\xi)\rangle = 0$, we have

$$J(x,\xi) = c(\xi) + \int_{\partial\Omega} b(x',\xi)G(x,x')\,dS_{x'},$$
$$\nabla_x J(x,\xi) = \int_{\partial\Omega} b(x',\xi)\nabla_x G(x,x')\,dS_{x'}. \tag{7.7}$$

Using the known fact that $|\nabla_x G(x,x')| \le C(\Omega)|x-x'|^{1-N}$, we then obtain from (7.6) and (7.7) that

$$|\nabla_x J(\xi,\xi)| \le C(\Omega)\int_{\partial\Omega}|\xi-x'|^{2-N}|x'-\xi|^{1-N}dS_{x'} \le C(\Omega)\mathrm{d}(\xi)^{2-N}.$$

This completes the proof. $\square$

8. Proof of Lemma 3.2. In this section, we prove Lemma 3.2.

We first consider the problem on $\mathbb{R}^N$, $N=2,3$. In what follows, $\langle f\rangle_* = \int_{\mathbb{R}^N} f(x)dx$.

Set $\sigma_0 = \langle W^2\rangle_*$. Then

$$\Delta W - W + \sigma_0^{-1}W^2 = 0 \qquad \text{in } \mathbb{R}^N\ .$$

Lemma 8.1. *Let* L_0 *be an operator defined as*

$$\mathrm{L}_0\phi \overset{\text{def}}{=} \Delta\phi - \phi + 2\sigma_0^{-1}W\phi.$$

Then the following conditions hold:

(1) *The principal eigenvalue* λ_0 *of* L_0 *is positive and its associated eigenfunction* ϕ_0 *is positive.*
(2) *Zero is an eigenvalue of* L_0 *with multiplicity* N*; its associated eigenspace is spanned by* $W_{x_1},\ldots,W_{x_N}$.
(3) *There exists* $\nu_0 > 0$ *such that*

$$\mathrm{L}_0(\phi,\phi) \overset{\text{def}}{=} \langle -|\nabla\phi|^2 - \phi^2 + 2\sigma_0^{-1}W\phi^2\rangle_* \ \le\ \nu_0\langle\phi^2\rangle_*$$

for all $\phi\in H^1(\mathbb{R}^N)$ *satisfying* $\phi\perp\phi_0, W_{x_1},\ldots,W_{x_N}$ *(in* $L^2(\mathbb{R}^N)$ *sense).*

This lemma follows directly from more general results of [16].

Lemma 8.2. *Let* L_1 *be an operator defined by*

$$\mathrm{L}_1\phi \overset{\text{def}}{=} \mathrm{L}_0\phi - \sigma_0^{-2}\langle W\phi\rangle_* W^2 - \sigma_0^{-2}\langle W^2\phi\rangle_* W. \tag{8.1}$$

Then L_1 *has the following properties:*

(1) *The operator* L_1 *is self-adjoint.*
(2) *The function* W *is an eigenfunction of* L_1 *with eigenvalue* $-\sigma_0^{-2}\langle W^3\rangle_*$.
(3) *For each* $i=1,\dots,N$, W_{x_i} *is an eigenfunction of* L_1 *with eigenvalue zero.*
(4) *Assume that* $N\le 3$. *Then there exists a positive constant* $\nu_1\in(0,1]$ *such that*

$$\mathrm{L}_1(\phi,\phi)\overset{\text{def}}{=}\langle-|\nabla\phi|^2-\phi^2\rangle_*+2\sigma_0^{-1}\langle W\phi^2\rangle_*-2\sigma_0^{-2}\langle W\phi\rangle_*\langle W^2\phi\rangle_*\;\le\;-\nu_1\langle\phi^2\rangle_*$$

for all $\phi\in H^1(\mathbb{R}^N)$ *satisfying* $\phi\perp W_{x_1},\dots,W_{x_N}$.

Proof. The first three assertions follow by direct verification. To prove assertion (4), we need to consider only those ϕ which are orthogonal to $W, W_{x_1},\cdots,W_{x_N}$. We will argue by contradiction. Since the essential spectrum of L_1 lies in $(-\infty,-1]$, if assertion (4) is not true, then there exists (λ,ϕ) such that

(i) λ is real and nonnegative,
(ii) $\phi\perp W, W_{x_1},\dots,W_{x_N}$, and
(iii) $\mathrm{L}_1\phi=\lambda\phi$.

We will show that conditions (i)–(iii) cannot hold simultaneously.

From the definition of L_1 and conditions (ii) and (iii), we have

$$(\mathrm{L}_0-\lambda)\phi=\sigma_0^{-2}\langle W^2\phi\rangle_*W. \tag{8.2}$$

First we claim that $\lambda\ne\lambda_0$. In fact, if $\lambda=\lambda_0$, then $(\mathrm{L}_0-\lambda_0)\phi\perp\phi_0$ so that $\langle W^2\phi\rangle_*\langle W\phi_0\rangle_*=0$. Consequently, as $\phi_0>0$ and $W>0$, $\langle W^2\phi\rangle_*=0$, so that $(\mathrm{L}_0-\lambda_0)\phi=0$. Thus ϕ is a multiple of ϕ_0. But this contradicts $\langle W^2\phi\rangle_*=0$. Hence, $\lambda\ne\lambda_0$.

Restricted to the space orthogonal to $W_{x_1},\dots,W_{x_N}$, $\mathrm{L}_0-\lambda$ is invertible, so that (8.2) implies that

$$\phi=\alpha(\mathrm{L}_0-\lambda)^{-1}W,\qquad \alpha=\sigma_0^{-2}\langle W^2\phi\rangle_*.$$

Hence, $\alpha\ne0$. Taking the inner product with $\sigma_0^{-2}W^2/\alpha$, we obtain

$$\begin{aligned}1&=\sigma_0^{-2}(W^2,(\mathrm{L}_0-\lambda)^{-1}W)\\&=\sigma_0^{-1}(\mathrm{L}_0W,(\mathrm{L}_0-\lambda)^{-1}W)\qquad(\text{as }\mathrm{L}_0W=\sigma_0^{-1}W^2)\\&=\sigma_0^{-1}((\mathrm{L}_0-\lambda)W,(\mathrm{L}_0-\lambda)^{-1}W)+\sigma_0^{-1}\lambda(W,(\mathrm{L}_0-\lambda)^{-1}W)\\&=\sigma_0^{-1}\langle W^2\rangle+\sigma_0^{-1}\lambda(W,(\mathrm{L}_0-\lambda)^{-1}W)\\&=1+\sigma_0^{-1}\lambda(W,(\mathrm{L}_0-\lambda)^{-1}W).\end{aligned}$$

Consider the function $F(z)=(W,(\mathrm{L}_0-z)^{-1}W)$ for $z\in(0,\lambda_0)\cup(\lambda_0,\infty)$. We have

$$F'(z)=(W,(\mathrm{L}_0-z)^{-2}W)=((\mathrm{L}_0-z)^{-1}W,(\mathrm{L}_0-z)^{-1}W)>0.$$

Since $\mathrm{L}_0(W+\frac12x\cdot\nabla W)=W$, ${\mathrm{L}_0}^{-1}W=W+\frac12x\cdot\nabla W+\Sigma_{i=1}^Nc_iW_{x_i}$. It then follows that $F(0)=(W,W+\frac12x\cdot\nabla W)=(1-\frac N4)\langle W^2\rangle>0$ as $N\le3$. Thus, $F(z)>0$ for all $z\in(0,\lambda_0)$. As $F(\infty)=0$, we also have $F(z)<0$ for all $z\in(\lambda_0,\infty)$. Hence, we have $\lambda\notin(0,\infty)$.

Finally, we show that $\lambda\ne0$. In fact, if $\lambda=0$, then as $\phi\perp W_{x_i}$ for all i, $\phi=\alpha{\mathrm{L}_0}^{-1}W=\alpha(W+\frac12x\cdot\nabla W)$. But this implies that $(\phi,W)=\alpha(1-\frac12)\langle W^2\rangle_*>0$, contradicting the assumption $\phi\perp W$. This completes the proof of the lemma. □

Remark 8.1. If $N = 4$, one sees that $\phi = W + \frac{1}{2}x \cdot \nabla W$ is an eigenfunction of L_0 with eigenvalue zero.

Next, we extend Lemma 8.2 to large balls $B_R = \{x \mid |x| < R\}$.

LEMMA 8.3. *Assume $N \leq 3$. There exist positive constants R_0 and ν_2 such that for each $R > R_0$ and each $\phi \in H^1(B_R)$ satisfying $\phi \perp W_{x_i}$, $i = 1, \ldots, N$ in $L^2(B_R)$ there holds*

$$\begin{aligned}(8.3)\quad {\mathrm{L}_1}^R(\phi,\phi) &\overset{\text{def}}{=} \int_{B_R}(-|\nabla\phi|^2 - \phi^2 + 2{\sigma_0}^{-1}W\phi^2) - 2\sigma_0^{-2}\int_{B_R} W\phi \int_{B_R} W^2\phi \\ &\leq\ -\nu_2 \int_{B_R}(|\nabla\phi|^2 + \phi^2).\end{aligned}$$

Proof. We will argue by contradiction. Suppose the assertion is not true. Then there exists a sequence $\{R_k, \phi_k\}_{k=1}^{\infty}$ such that $R_k > k, \phi_k \in H^1(B_{R_k})$, $\phi_k \perp W_{x_i}$ in $L^2(B_{R_k})$ for all i, $\int_{B_{R_k}}(|\nabla\phi_k|^2 + \phi_k^2) = 1$, and

$$(8.4)\qquad \limsup_{k\to\infty} {\mathrm{L}_1}^{R_k}(\phi_k,\phi_k) \geq 0.$$

Since H^1 is weakly compact and the embedding $H^1 \to L^2$ is compact, we can assume, by taking a subsequence if necessary, that there exists $\phi \in H^1(\mathbb{R}^N)$ such that $\lim_{k\to\infty}\phi_k = \phi$, weakly in $H^1(B_R)$ and strongly in $L^2(B_R)$ for every $R > 0$. In addition, $\|\phi\|_{H^1(\mathbb{R}^N)} \leq 1$.

Since W decays exponentially fast, we have, $\langle\phi W_{x_i}\rangle_* = \lim_{k\to\infty}\int_{B_{R_k}} \phi_k W_{x_i} = 0$ for all $i = 1, \ldots, N$. In addition, as $k \to \infty$,

$$\begin{aligned}\gamma_k &\overset{\text{def}}{=} \int_{B_k} 2\sigma_0 {-} 1 W\phi_k - 2\sigma_0^{-2}\int_{B_{R_k}} W\phi_k \int_{B_{R_k}} W^2\phi_k \\ &\to \gamma \overset{\text{def}}{=} 2\sigma_0^{-1}\langle W\phi^2\rangle_* - 2\sigma_0^{-2}\langle W\phi\rangle_*\langle W^2\phi\rangle_*.\end{aligned}$$

If $\phi \equiv 0$, then $\gamma = 0$ so that

$${\mathrm{L}_1}^{R_k}(\phi_k,\phi_k) = \gamma_k - 1 < -1/2$$

for all large k. But this contradicts (8.4).

If $\phi \not\equiv 0$, then by Lemma 8.2, as $\phi \perp W_{x_j}$ for all j, $\mathrm{L}_1(\phi,\phi) \leq -\nu_1\langle\phi^2\rangle_*$. As $\phi \in H^1(R)$, there exists a large M such that $\|\phi\|^2_{H^1(\mathbb{R}^N\setminus B_M)} \leq \frac{1}{2}\nu_1\langle\phi^2\rangle_*$. Consequently,

$$\gamma - \int_{B_M}(|\nabla\phi|^2 + \phi^2)dx \leq -\frac{1}{2}\nu_1\langle\phi^2\rangle_*.$$

It then follows that

$$\limsup_{k\to\infty} {\mathrm{L}_1}^{R_k}(\phi_k,\phi_k) \leq \limsup_{k\to\infty}\left\{\gamma_k - \int_{B_M}(|\nabla\phi_k|^2 + \phi_k^2)\right\} \leq \gamma - \int_{B_M}(|\nabla\phi|^2 + \phi^2) < 0\ .$$

Again, we obtain a contradiction. The proof is now complete. □

Proof of Lemma 3.2. We will denote

$$\langle f\rangle_\varepsilon = \int_{\Omega_\varepsilon} f, \qquad \Omega_\varepsilon = \frac{1}{\varepsilon}\Omega.$$

We also set

$$\mathcal{L}^{\xi,\varepsilon}(\phi,\psi) \overset{\text{def}}{=} \langle -\nabla\phi\cdot\nabla\psi - \phi\psi + 2\sigma^{-1}w\phi\psi\rangle_\varepsilon - 2\sigma^{-2}\langle w\phi\rangle_\varepsilon\langle w^2\psi\rangle_\varepsilon.$$

We will show that there exists a positive constant ν which is independent of ε such that for every sufficiently small positive ε,

$$\mathcal{L}^{\xi,\varepsilon}(\phi,\phi) \le -\nu\langle|\nabla\phi|^2+\phi^2\rangle_\varepsilon \quad \forall\,\phi, \langle\phi, w_{\xi_i}(\cdot,\xi)\rangle_\varepsilon = 0, \tag{8.5}$$

which is equivalent to the spectral estimate in Lemma 3.2. By scaling, we can assume that $\langle\phi^2+|\nabla\phi|^2\rangle_\varepsilon = 1$. Set $\mathrm{d}_\varepsilon(\xi) = \frac{\mathrm{dist}(\xi)}{\varepsilon}$ and $R = \mathrm{d}_\varepsilon(\xi)$. (Note that $R \ge 2|\ln(D\varepsilon^{-2})| \ge R_0$ if ε is sufficiently small.) Translating Ω_ε if necessary, we can always achieve $\xi = 0$.

Let ${\mathrm{L}_1}^R(\phi,\phi)$ be defined as in (8.4). As $|w-W|_{L^\infty} + |\sigma-\sigma_0| = O(\varepsilon)$ and $|W| = O(\varepsilon)$ outside B_R, we have

$$\mathcal{L}^{\xi,\varepsilon}(\phi,\phi) = L^R(\phi,\phi) - \int_{\Omega_\varepsilon\setminus B_R}(|\nabla\phi|^2+\phi^2) + O(\varepsilon). \tag{8.6}$$

Now let $\phi^R = \phi - \sum_{i=1}^N c_iW_{x_i}$ be the $L^2(B_R)$ orthogonal projection of ϕ on $\{W_{x_1},\dots,W_{x_N}\}^\perp$. Then, $0 = \langle\phi W_{x_i}\rangle_\varepsilon = \int_{B_R}\phi W_{x_i} + O(\varepsilon)$, so that $c_i = O(\varepsilon)$ for all i. Hence, $\|\phi-\phi^R\|_{H^1(B_R)} = O(\varepsilon)$ and

$${\mathrm{L}_1}^R(\phi,\phi) = {\mathrm{L}_1}^R(\phi^R,\phi^R) + O(\varepsilon). \tag{8.7}$$

From Lemma 8.2 we have

$${\mathrm{L}_1}^R(\phi^R,\phi^R) \le -\nu_1\int_{B_R}\Big(|\nabla\phi^R|^2 + (\phi^R)^2\Big) = -\nu_1\int_{B_R}(|\nabla\phi|^2+\phi^2) + O(\varepsilon)\ .$$

Combining this with (8.6) and (8.7) then gives

$$\begin{aligned}\mathcal{L}^{\xi,\varepsilon}(\phi,\phi) &\le -\int_{\Omega_\varepsilon\setminus B_R}(|\nabla\phi|^2+\phi^2) - \nu_1\int_{B_R}(|\nabla\phi|^2+\phi^2) + O(\varepsilon)\\ &\le -\nu_1\langle\phi^2+|\nabla\phi|^2\rangle_\varepsilon + O(\varepsilon).\end{aligned} \tag{8.8}$$

Taking ε sufficiently small, we then obtain the assertion of the lemma. □

REFERENCES

[1] N. D. Alikakos and G. Fusco, *Slow dynamics for the Cahn-Hilliard equation in higher space dimensions,* I: *Spectral estimates*, Comm. Partial Differential Equations, 19 (1994), pp. 1397–1447.

[2] N. D. Alikakos and G. Fusco, *Slow dynamics for the Cahn-Hilliard equation in higher space dimensions,* II: *The motion of bubbles*, Arch. Ration. Mech. Anal, 141 (1998), pp. 1–61.

[3] X. Chen and M. Kowalczyk, *Slow dynamics of interior spikes in the shadow Gierer–Meinhardt system*, Adv. Differential Equations, 6 (2001), pp. 847–872.

[4] C. V. Coffman, *Uniqueness of the ground state solution for* $\Delta u - u + u^3 = 0$ *and a variational characterization of other solutions*, Arch. Ration. Mech. Anal., 46 (1972), pp. 81–95.

[5] E. N. Fraenkel, *An Introduction to Maximum Principle and Symmetry in Elliptic Problems*, Cambridge University Press, Cambridge, UK, 1999.

[6] A. Gierer and H. Meinhardt, *A theory of biological pattern formation*, Kybernetik, 12 (1972), pp. 30–39.

[7] D. Iron and M. J. Ward, *A metastable spike solution for a nonlocal reaction-diffusion model*, SIAM J. Appl. Math., 60 (2000), pp. 778–802.

[8] D. Iron, M. J. Ward, and J. Wei, *The stability of spike solutions to the one-dimensional Gierer–Meinhardt model*, Phys. D, to appear.

[9] M. K. Kwang, *Uniqueness of positive solutions of $\Delta u - u + u^p = 0$ in $\mathbb{R}^N$*, Arch. Ration. Mech. Anal., 105 (1989), pp. 243–266.

[10] K. McLeod & J. Serrin, *Uniqueness of positive radial solutions of $\Delta u + f(u) = 0$ in $\mathbb{R}^N$*, Arch. Ration. Mech. Anal., 99 (1987), pp. 115–145.

[11] C.-S. Lin, W.-M. Ni, and I. Takagi, *Large amplitude stationary solutions to a chemotaxis system*, J. Differential Equations, 72 (1988), pp. 1–27.

[12] H. Meinhardt, *Models of Biological Pattern Formation*, Academic Press, London, 1982.

[13] W.-M. Ni, *Diffusion, cross-diffusion, and their spike-layer steady states*, Notices Amer. Math. Soc., 45 (1998), pp. 9–18.

[14] W.-M. Ni and I. Takagi, *On the Neumann problem for some semilinear elliptic equations and systems of activator-inhibitor type*, Trans. Amer. Math. Soc., 297 (1986), pp. 351–368.

[15] W.-M. Ni and I. Takagi, *On the shape of least-energy solutions to a semilinear Neumann problem*, Comm. Pure Appl. Math., 44 (1991), pp. 819–851.

[16] 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.

[17] W.-M. Ni, I. Takagi, and E. Yanagida, *Stability analysis of point condensation solutions to a reaction-diffusion system*, Tohoku Math. J. (2), to appear.

[18] L. A. Peletier and J. Serrin, *Uniqueness of positive solutions of semilinear equations in $\mathbb{R}^N$*, J. Differential Equations, 61 (1983), pp. 181–197.

[19] J. Wei, *On single interior spike solutions of Gierer–Meinhardt system: Uniqueness and spectrum estimates*, European J. Appl. Math., 10 (1999), pp. 353–378.

[20] J. Wei and M. Winter, *Spikes for the two-dimensional Gierer–Meinhardt system: Strong coupling*, J. Differential Equations, to appear.

[21] L. Zhang, *Uniqueness of ground state solutions*, Acta Math. Sci. (English Ed.), 6 (1988), pp. 449–468.

SIAM J. Math. Anal.
Vol. 33, No. 1, pp. 194–215

BIORTHOGONAL WAVELET SPACE: PARAMETRIZATION AND FACTORIZATION[*]

HOWARD L. RESNIKOFF[†], JUN TIAN[‡¶], AND RAYMOND O. WELLS, JR.[§¶]

Abstract. In this paper we study the algebraic and geometric structure of the space of compactly supported biorthogonal wavelets. We prove that any biorthogonal wavelet matrix pair (which consists of the scaling filters and wavelet filters) can be factored as the product of primitive paraunitary matrices, a pseudo identity matrix pair, an invertible matrix, and the canonical Haar matrix. Compared with the factorization results of orthogonal wavelets, it now becomes apparent that the difference between orthogonal and biorthogonal wavelets lies in the pseudo identity matrix pair and the invertible matrix, which in the orthogonal setting will be the identity matrix and a unitary matrix. Thus by setting the pseudo identity matrix pair to be the identity matrix and using the Schmidt orthogonalization method on the invertible matrix, it is very straightforward to convert a biorthogonal wavelet pair into an orthogonal wavelet.

Key words. biorthogonal wavelets, parametrization, Pollen product, vanishing moment

AMS subject classifications. 42C15, 15A23

PII. S0036141099360361

1. Introduction.

The theory of wavelet analysis has grown explosively in the last fifteen years. The terminology "wavelet" was first introduced, in the context of a mathematical transform, in 1984 by Grossmann and Morlet [12]. In 1988, Daubechies [9] introduced a class of compactly supported orthogonal wavelets with growing smoothness for increasing support. Mallat [22] and Meyer [23] presented the theory of multiresolution analysis. The spline family was first studied by Battle [1], Chui [5], and Lemarié [21]. The necessary and sufficient conditions for orthonormal wavelet bases were given by Cohen [7] and Lawton [20]. Except for the *Haar wavelet*, compactly supported orthogonal wavelets can not be symmetric, though symmetry is highly desired, for example, in the applications in signal processing, where symmetry corresponds to linear phase. To obtain symmetry and keep the property of perfect reconstruction, the orthogonal condition was replaced by biorthogonality and the theory of biorthogonal wavelets [6, 8, 30] was established. (One of the earliest examples of a nonorthogonal biorthogonal representation is in Paley and Wiener's book on the Fourier transform in the complex domain [24].) At the same time, pioneer work has been done by many scientists from mathematics, physics, and engineering. For more details of wavelet theory, we refer to [3, 4, 10, 14, 23, 26, 27, 31, 32].

In this paper we will study the algebraic and geometric structure of the space of compactly supported biorthogonal wavelets. We will prove that any biorthogonal wavelet matrix pair can be decomposed into four components: an orthogonal component $V(z)$, a pseudo identity matrix pair, an invertible matrix G, and a constant matrix $\mathbf{H}$. There have been several factorization results for biorthogonal wavelets

[*] Received by the editors August 19, 1999; accepted for publication (in revised form) October 5, 2000; published electronically April 24, 2001.

http://www.siam.org/journals/sima/33-1/36036.html

[†] FutureWAVE, Inc., 85 East India Row #35 G, Boston, MA 02110 (howard@resnikoff.com).

[‡] Digimarc Corporation, 19801 SW 72nd Avenue, Tualatin, OR 97062 (jtian@digimarc.com).

[§] International University Bremen, Postfach 750561, 28725 Bremen, Germany (wells@iu-bremen.de).

[¶] The work of this author was supported in part by AFOSR/DARPA, Exxon Production Research, NSF, Northrop Grumman Corporation, and Texas ATP.

reported in [2, 11, 14, 19], etc. Our contribution in this paper is that we require a zeroth order vanishing moment condition (2.3) on the wavelet matrix. This vanishing moment condition is necessary, because it is required for the existence of scaling functions and wavelet functions. So we are placing emphasis on the wavelet factorization instead of perfect reconstruction filter banks. This paper is also partly motivated by the problem of finding nontrivial mappings between orthogonal and biorthogonal wavelets. To achieve this goal, we arrange our factorization formula so that it is closely connected with the factorization formula for orthogonal wavelets, which is another difference between our factorization and those in [2, 11, 14, 19].

The paper is organized as follows. In section 2 we give a brief review of some definitions and properties of biorthogonal and orthogonal wavelets. The parametrization of biorthogonal wavelets is presented in section 3. We will define a group structure on biorthogonal wavelets under the Pollen product. Because of its special role in the biorthogonal factorization, the pseudo identity matrix pair will be studied in section 4. We propose a sufficient condition on the pseudo identity matrix pair such that it can be factored as the product of primitive pseudo identity matrices. We discuss conversion between orthogonal and biorthogonal wavelets in section 5, and a simple example is given in section 6. We conclude the paper in section 7.

Note that in this paper we will assume entries in wavelet matrices are real-valued. The generalization to a subfield F of complex numbers $\mathbf{C}$ *closed under complex conjugation* is straightforward. Some examples of F will be the rational numbers $\mathbf{Q}$, the real numbers $\mathbf{R}$, the Gaussian rational numbers $\mathbf{Q}(i) := \mathbf{Q} + i\mathbf{Q}$, and the complex numbers $\mathbf{C}$ itself.

2. Preliminaries. For a matrix $A = (a_{i,j})$ consisting of m rows of vectors with only a finite number of entries $a_{i,j}$ being nonzero, define submatrices A_k of size $m \times m$ of A in the following manner:

$$A_k = (a_{i,km+j}), \quad 0 \le i \le m-1, 0 \le j \le m-1.$$

In other words, A is expressed in terms of block matrices in the form

$$A = (\ldots, A_{-1}, A_0, A_1, \ldots),$$

where, for instance,

$$A_0 = \begin{pmatrix} a_{0,0} & a_{0,1} & \cdots & a_{0,m-1} \\ \vdots & & & \vdots \\ a_{m-1,0} & a_{m-1,1} & \cdots & a_{m-1,m-1} \end{pmatrix}.$$

From the matrix A, we construct the formal power series

$$A(z) := \sum_{k \in \mathbf{Z}} A_k z^{-k} = \sum_{k=k_0}^{k_1} A_k z^{-k}, \tag{2.1}$$

where k_0 and k_1 are the smallest and largest indices that $A_k \ne 0$, respectively. We call $A(z)$ the *Laurent series* of the matrix A and k_0 the *leading index*. We can equally well write $A(z)$ as an $m \times m$ matrix

$$A(z) = \begin{pmatrix} \sum_k a_{0,km} z^{-k} & \cdots & \sum_k a_{0,km+m-1} z^{-k} \\ \vdots & & \vdots \\ \cdots & \sum_k a_{i,km+j} z^{-k} & \cdots \\ \vdots & & \vdots \\ \sum_k a_{m-1,km} z^{-k} & \cdots & \sum_k a_{m-1,km+m-1} z^{-k} \end{pmatrix},$$

which we will refer to as the *polyphase decomposition* of A. For the case of $m = 2$, we find

$$A(z) = \begin{pmatrix} \cdots a_{0,-2} z + a_{0,0} + a_{0,2} z^{-1} \cdots, & \cdots a_{0,-1} z + a_{0,1} + a_{0,3} z^{-1} \cdots \\ \cdots a_{1,-2} z + a_{1,0} + a_{1,2} z^{-1} \cdots, & \cdots a_{1,-1} z + a_{1,1} + a_{1,3} z^{-1} \cdots \end{pmatrix}.$$

Let

$$g := k_1 - k_0 + 1$$

be the number of nonzero terms in the summation (2.1) and call g the *genus* of the Laurent series $A(z)$ and the matrix A. Thus A has a size of $m \times mg$. Finally we define the *adjoint* $\tilde{A}(z)$ of the Laurent series $A(z)$ by

$$\tilde{A}(z) := A^*(z^{-1}) := \sum_{k=k_0}^{k_0+g-1} A_k^* z^k = \sum_{k=-k_0-g+1}^{-k_0} A_{-k}^* z^{-k},$$

where $A_k^* := \overline{A_k^t}$ is the Hermitian adjoint of the $m \times m$ matrix A_k. When A_k is a real matrix, $A_k^* = A_k^t$.

DEFINITION 2.1. *A pair of $m \times mg$ matrices $L = (l_{i,j}), R = (r_{i,j})$ is said to be a* biorthogonal wavelet matrix pair *of rank m and genus g if their polyphase decompositions $L(z)$ and $R(z)$ satisfy*

$$L(z)\tilde{R}(z) = mI_m \tag{2.2}$$

and

$$\sum_j l_{i,j} = \sum_j r_{i,j} = \begin{cases} m & \text{if } i = 0, \\ 0 & \text{if } 1 \le i \le m-1, \end{cases} \tag{2.3}$$

where I_m is the $m \times m$ identity matrix.

We call L the *analysis matrix* and R the *synthesis matrix* of the biorthogonal wavelet matrix pair. We will refer to (2.2) and (2.3) as the *perfect reconstruction* and *linear* conditions defining a biorthogonal wavelet matrix pair, respectively. The perfect reconstruction condition (2.2) is equivalent to saying that $L(z), R(z)$ are invertible polynomials in $\mathrm{SL}(m; \mathbf{R}[z, z^{-1}])$ (a proof can be found in [11]), that is,

$$\det(L(z)) = c_l z^{-b}, \quad \det(R(z)) = c_r z^{-b}, \quad \text{with } c_l c_r = m \tag{2.4}$$

for some integer b. Note that in the theory of wavelet analysis, we will systematically employ the linear constraint (2.3) in addition to the perfect reconstruction condition (2.2). Actually (2.3) (which is exactly the zeroth order vanishing moment condition of scaling functions and wavelet functions) is a necessary condition for the existence of scaling functions and wavelet functions. This is one of the main differences between wavelets and perfect reconstruction filter banks.

For a biorthogonal wavelet matrix pair (L, R), suppose that $\phi(x)$ and $\tilde{\phi}(x)$ are L^2 solutions of the refinement equations

$$\phi(x) = \sum_k l_{0,k}\phi(mx-k), \quad \tilde{\phi}(x) = \sum_k r_{0,k}\tilde{\phi}(mx-k),$$

and define $\psi^1(x), \dots, \psi^{m-1}(x), \tilde{\psi}^1(x), \dots, \tilde{\psi}^{m-1}(x)$ by

$$\psi^i(x) := \sum_k r_{i,k}\phi(mx-k), \quad \tilde{\psi}^i(x) := \sum_k l_{i,k}\tilde{\phi}(mx-k), \quad i = 1, \dots, m-1,$$

then $\psi^1_{j,k}, \dots, \psi^{m-1}_{j,k}, \tilde{\psi}^1_{j,k}, \dots, \tilde{\psi}^{m-1}_{j,k}$ constitute a weak dual frame of $L^2(\mathbf{R})$, that is, for any $f(x), g(x) \in L^2(\mathbf{R})$,

$$\lim_{J\to\infty} \sum_{i=1}^{m-1} \sum_{j=-J}^{J} \sum_{k=-\infty}^{\infty} \langle f, \psi^i_{j,k}\rangle \langle \tilde{\psi}^i_{j,k}, g\rangle = \langle f, g\rangle,$$

where

$$\psi^i_{j,k}(x) := m^{j/2}\psi^i(m^j x - k), \quad \tilde{\psi}^i_{j,k}(x) := m^{j/2}\tilde{\psi}^i(m^j x - k).$$

We call $\phi(x), \psi^1(x), \dots, \psi^{m-1}(x), \tilde{\phi}(x), \tilde{\psi}^1(x), \dots, \tilde{\psi}^{m-1}(x)$ the *analysis scaling function*, *analysis wavelet functions*, *synthesis scaling function*, and *synthesis wavelet functions*, respectively. With some additional conditions [6, 8, 20, 30], one can derive a (strong) dual frame and even dual Riesz bases of $L^2(\mathbf{R})$.

A special and also widely used subset of biorthogonal wavelets are the orthogonal wavelets.

DEFINITION 2.2. *An $m \times mg$ matrix A is said to be an* orthogonal wavelet matrix *of rank m and genus g if (A, A) is a biorthogonal wavelet matrix pair.*

Similarly, starting from an orthogonal wavelet matrix A, one can define the *scaling function* $\phi(x)$ and the *wavelet functions* $\psi^1(x), \dots, \psi^{m-1}(x)$ and their rescaled and translated version $\psi^i_{j,k}$. In this setting, the wavelet functions $\psi^1(x), \dots, \psi^{m-1}(x)$ will generate a tight frame, that is, for any $f(x) \in L^2(\mathbf{R})$,

$$f(x) = \sum_{i=1}^{m-1} \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} (d_i)_{j,k}\psi^i_{j,k}(x),$$

where the coefficients are given by

$$(d_i)_{j,k} = \int_{\mathbf{R}} f(x)\psi^i_{j,k}(x)\,dx.$$

With an additional Cohen condition [7] or equivalent Lawton condition [20], $\psi^i_{j,k}(x)$ constitute an orthonormal basis of $L^2(\mathbf{R})$.

3. Parametrization of biorthogonal wavelets. In this section we formulate the parametrization of the space of compactly supported biorthogonal wavelets. Borrowing the eigenfilter approach from [29], the following lemma will be derived.

LEMMA 3.1. *If (L, R) is a biorthogonal wavelet matrix pair of rank m and genus g, then there exist unit column vectors $v_1, v_2, \ldots, v_d$ such that*

$$L(z) = z^{-k_0} V_1(z) V_2(z) \cdots V_d(z) C(z) L(1), \tag{3.1}$$

$$R(z) = z^{-k_0} V_1(z) V_2(z) \cdots V_d(z) D(z) R(1), \tag{3.2}$$

where k_0 is the leading index of L, $d = b - mk_0$, where b is the integer appearing in (2.4)*, and*

$$V_i(z) = I_m - v_i v_i^* + v_i v_i^* z^{-1}, \quad i = 1, 2, \ldots, d,$$

$$C(z) = \sum_{k=0}^{k_c} C_k z^{-k}, \quad D(z) = \sum_{k=k_d}^{0} D_k z^{-k},$$

with $k_c \leq g - 1$, and

$$C(z)\tilde{D}(z) = I_m, \quad \det(C(z)) = \det(D(z)) = 1, \quad C(1) = D(1) = I_m.$$

Let v be a unit column vector, that is, $v^* v = 1$. Define the Laurent matrix

$$V(z) := I - vv^* + vv^* z^{-1}, \tag{3.3}$$

and let V be the corresponding $m \times 2m$ matrix. We will say that a matrix V of the form (3.3) is a *primitive paraunitary matrix*. It is easy to verify (see [29]) that

$$V(z)\tilde{V}(z) = I_m, \quad \det(V(z)) = z^{-1}.$$

Proof of Lemma 3.1. Without loss of generality, assume $k_0 = 0$ (and so $b = d$), and

$$L(z) = \sum_{k=0}^{g-1} L_k z^{-k}, \quad R(z) = \sum_{k=k_r}^{k_r+g-1} R_k z^{-k}.$$

Now

$$L(z)\tilde{R}(z) = \left(\sum_{k=0}^{g-1} L_k z^{-k}\right)\left(\sum_{k=k_r}^{k_r+g-1} R_k^* z^k\right) = \left(\sum_{k=0}^{g-1} L_k z^{-k}\right)\left(\sum_{k=-k_r-g+1}^{-k_r} R_{-k}^* z^{-k}\right)$$
$$= L_0 R_{k_r+g-1}^* z^{k_r+g-1} + \cdots + L_{g-1} R_{k_r}^* z^{k_r-g+1} = mI_m.$$

From (2.4)

$$\det(L(z)) = c_l z^{-d}, \quad \det(R(z)) = c_r z^{-d}, \quad \text{with } c_l c_r = m$$

for some integer $d \geq 0$.

If $d > 0$, then $k_r + g - 1$ must be positive (otherwise $\det(R(z)) \neq c_r z^{-d}$). It follows that

$$L_0 R_{k_r+g-1}^* = 0.$$

So L_0 must be a singular matrix (if $R^*_{k_r+g-1} = 0$, then $L_0 R^*_{k_r+g-2} = 0$). Then choose a vector v_1 of unit length such that

$$v_1{}^* L_0 = 0,$$

and define

$$V_1(z) = I_m - v_1 v_1{}^* + v_1 v_1{}^* z^{-1},$$

as a primitive paraunitary matrix. Note that

$$\tilde{V}_1(z) = I_m - v_1 v_1{}^* + v_1 v_1{}^* z.$$

Define

$$L^1(z) = \tilde{V}_1(z) L(z), \quad R^1(z) = \tilde{V}_1(z) R(z).$$

It follows that (L^1, R^1) is a biorthogonal wavelet matrix pair of rank m and

$$L(z) = V_1(z) L^1(z), \quad R(z) = V_1(z) R^1(z). \tag{3.4}$$

Notice that

$$\begin{aligned} L^1(z) &= (I_m - v_1 v_1{}^* + v_1 v_1{}^* z) \left(\sum_{k=0}^{g-1} L_k z^{-k} \right) \\ &= \left(v_1 v_1{}^* L_1 + (I_m - v_1 v_1{}^*) L_0\right) + \cdots + (I_m - v_1 v_1{}^*) L_{g-1} z^{-g+1} = \sum_{k=0}^{l^1} L^1_k z^{-k}, \end{aligned}$$

where $l^1 \le g-1$. Take the determinant on both sides of (3.4)

$$c_l z^{-d} = z^{-1} \det(L^1(z)), \quad c_r z^{-d} = z^{-1} \det(R^1(z)),$$

which implies that

$$\det(L^1(z)) = c_l z^{-d+1}, \quad \det(R^1(z)) = c_r z^{-d+1}.$$

Thus the degree of the determinant $\det(L^1(z))$ is increased by 1, compared with $\det(L(z))$. Proceeding in this fashion, we obtain

$$L(z) = V_1(z) V_2(z) \cdots V_d(z) L^d(z), \quad R(z) = V_1(z) V_2(z) \cdots V_d(z) R^d(z), \tag{3.5}$$

where $\det(L^d(z)) = c_l, \det(R^d(z)) = c_r$, and

$$L^d(z) = \sum_{k=0}^{l^d} L^d_k z^{-k}, \quad R^d(z) = \sum_{k=k_1}^{k_2} R^d_k z^{-k},$$

with $l^d \le l^{d-1} \le \cdots \le l^1 \le g-1$. We claim that $k_2 = 0$. If not, then k_2 must be positive (since $\det(R^d(z)) = c_r$). Apply the factorization one more time to obtain $L^d(z) = V_{d+1}(z) L^{d+1}(z)$, where $\det(L^{d+1}(z)) = c_l z$ which is impossible, since $L^{d+1}(z)$ is the summation of nonpositive powers of z.

Because $L(1)\tilde{R}(1) = mI_m$, $L(1), R(1)$ are invertible matrices. Now set

$$L^d(z) = C(z) L(1), \quad R^d(z) = D(z) R(1), \tag{3.6}$$

with

$$C(z) = \sum_{k=0}^{k_c} C_k z^{-k}, \quad D(z) = \sum_{k=k_d}^{0} D_k z^{-k},$$

where $k_c = l^d, C_k = L^d_k(L(1))^{-1}, k_d = k_1, D_k = R^d_k(R(1))^{-1}$. Since $V_i(1) = I_m, i = 1, \ldots, d$, we have $\det(C(z)) = \det(D(z)) = 1$. Combining (3.5) and (3.6), the lemma follows. ◻

Remark 1. It is possible to have $d = 0$ and $C(z) \neq I_m$, that is, a biorthogonal wavelet matrix pair with no primitive paraunitary matrices in the factorization. A simple example is

$$L(z) = \begin{pmatrix} 2 - z^{-1} & z^{-1} \\ -1 & 1 \end{pmatrix}, \quad R(z) = \begin{pmatrix} 1 & 1 \\ -z & -z+2 \end{pmatrix}. \tag{3.7}$$

Next we will study the structure of the matrix pair $(L(1), R(1))$ appearing in the factorization of Lemma 3.1. For this purpose, we define a *biorthogonal Haar wavelet matrix pair* (H_L, H_R) to be a biorthogonal wavelet matrix pair with genus g equal to 1. Thus in the polyphase decomposition there is exactly only one term in the summations, that is,

$$H_L(z) = z^{-k_L} \begin{pmatrix} l_{0,0} & \cdots & l_{0,m-1} \\ \vdots & & \vdots \\ \cdots & l_{i,j} & \cdots \\ \vdots & & \vdots \\ l_{m-1,0} & \cdots & l_{m-1,m-1} \end{pmatrix}$$

and

$$H_R(z) = z^{-k_R} \begin{pmatrix} r_{0,0} & \cdots & r_{0,m-1} \\ \vdots & & \vdots \\ \cdots & r_{i,j} & \cdots \\ \vdots & & \vdots \\ r_{m-1,0} & \cdots & r_{m-1,m-1} \end{pmatrix},$$

where $k_L, k_R \in \mathbf{Z}$. It is easy to verify that $k_L = k_R$. And without loss of generality, we will always assume $k_L = k_R = 0$ for a biorthogonal Haar wavelet matrix pair (H_L, H_R), since a multiplication with z^{-k} is just a shift of index j in $l_{i,j}$ and $r_{i,j}$.

Let us now provide a characterization of biorthogonal Haar wavelet matrix pairs. Recall that the general linear group GL_{m-1} is the group of $(m-1) \times (m-1)$ matrices G such that G is invertible.

Theorem 3.2. *Two $m \times m$ matrices H_L, H_R constitute a biorthogonal Haar wavelet matrix pair if and only if*

$$H_L = \begin{pmatrix} 1 & 0 \\ 0 & G \end{pmatrix} \mathbf{H}, \quad H_R = \begin{pmatrix} 1 & 0 \\ 0 & (G^{-1})^* \end{pmatrix} \mathbf{H},$$

where $G \in \mathrm{GL}_{m-1}$ is an invertible matrix, and $\mathbf{H}$ is the canonical Haar matrix *of* *rank m, which is defined by*

$$\mathbf{H} := \begin{pmatrix} 1 & 1 & \cdots & \cdots & \cdots & \cdots & 1 \\ -\sqrt{m-1} & \sqrt{\frac{1}{m-1}} & \cdots & \cdots & \cdots & \cdots & \sqrt{\frac{1}{m-1}} \\ \vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & -\sqrt{\frac{im}{i+1}} & \sqrt{\frac{m}{i^2+i}} & \cdots & \sqrt{\frac{m}{i^2+i}} \\ \vdots & \cdots & \cdots & \cdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & -\sqrt{\frac{m}{2}} & \sqrt{\frac{m}{2}} \end{pmatrix},$$

where $i = m-1, \dots, 2, 1$.

Proof. First, assume H_L and H_R are of the form

$$H_L = \begin{pmatrix} 1 & 0 \\ 0 & G \end{pmatrix} \mathbf{H}, \quad H_R = \begin{pmatrix} 1 & 0 \\ 0 & (G^{-1})^* \end{pmatrix} \mathbf{H}.$$

By the definition of the canonical Haar matrix $\mathbf{H}$, we have

$$\mathbf{H}\tilde{\mathbf{H}} = mI_m, \quad \mathbf{H} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} m \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Now it is easy to check that H_L and H_R will satisfy (2.2) and (2.3). Thus H_L and H_R is a biorthogonal Haar wavelet matrix pair.

Conversely, if (H_L, H_R) is a biorthogonal Haar wavelet matrix pair, define

$$L = \frac{1}{m} H_L \tilde{\mathbf{H}}, \quad R = \frac{1}{m} H_R \tilde{\mathbf{H}},$$

then

$$H_L = L\mathbf{H}, \quad H_R = R\mathbf{H}.$$

From the linear constraint (2.3)

$$\begin{pmatrix} m \\ 0 \\ \vdots \\ 0 \end{pmatrix} = H_L \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = L\mathbf{H} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = L \begin{pmatrix} m \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

then L should be of the form

$$L = \begin{pmatrix} 1 & \alpha \\ 0 & G_1 \end{pmatrix},$$

where α is an $(m-1)$ dimensional row vector and G_1 is an $(m-1) \times (m-1)$ matrix. Similarly

$$R = \begin{pmatrix} 1 & \beta \\ 0 & G_2 \end{pmatrix}$$

for an $(m-1)$ dimensional row vector β and an $(m-1)\times(m-1)$ matrix G_2. Since

$$mI_m = H_L\tilde{H}_R = L\mathbf{H}\tilde{\mathbf{H}}\tilde{R} = mL\tilde{R},$$

it follows that

$$L\tilde{R} = I_m,$$

that is,

$$\begin{pmatrix} 1 & \alpha \\ 0 & G_1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ \beta^* & {G_2}^* \end{pmatrix} = I_m.$$

Thus

$$G_2 = ({G_1}^{-1})^*, \quad \alpha = \beta = 0. \qquad \square$$

Let (L, R) be a biorthogonal wavelet matrix pair of rank m and let $L(z), R(z)$ be their Laurent series. Define the *characteristic Haar matrix pair* $(\chi(L), \chi(R))$ of (L, R) by

$$\chi(L) := L(1), \quad \chi(R) := R(1).$$

We now have the following theorem which relates biorthogonal wavelet matrix pairs to biorthogonal Haar wavelet matrix pairs.

THEOREM 3.3. *If (L, R) is a biorthogonal wavelet matrix pair of rank m, then $(\chi(L), \chi(R))$ is a biorthogonal Haar wavelet matrix pair of rank m. Thus χ is a well defined mapping from biorthogonal wavelet matrix pairs of rank m to biorthogonal Haar wavelet matrix pairs of rank m.*

Proof. Evaluate (2.2) and (2.3) at $z = 1$; thus it follows that $(\chi(L), \chi(R))$ is a biorthogonal wavelet matrix pair of rank m and genus 1. $\square$

With Lemma 3.1, Theorem 3.2, and Theorem 3.3, we now have a complete characterization of biorthogonal wavelet matrix pairs.

THEOREM 3.4 (biorthogonal factorization theorem). *A pair of $m \times mg$ matrices (L, R) is a biorthogonal wavelet matrix pair of rank m if and only if there exist primitive paraunitary matrices $V_1, \dots, V_d$, $d \geq 0$, such that*

$$L(z) = z^{-k_0}V_1(z)V_2(z)\cdots V_d(z)C(z)\begin{pmatrix} 1 & 0 \\ 0 & G \end{pmatrix}\mathbf{H},$$

$$R(z) = z^{-k_0}V_1(z)V_2(z)\cdots V_d(z)D(z)\begin{pmatrix} 1 & 0 \\ 0 & (G^{-1})^* \end{pmatrix}\mathbf{H},$$

where $k_0 \in \mathbf{Z}$, $d = b - mk_0$, b is the exponent of $\det(L(z))$, *$G \in \mathrm{GL}_{m-1}$ is an invertible matrix, $\mathbf{H}$ is the canonical Haar matrix of rank m, and*

$$C(z) = \sum_{k=0}^{k_c} C_k z^{-k}, \quad D(z) = \sum_{k=k_d}^{0} D_k z^{-k},$$

with $k_c \leq g-1$, and

$$C(z)\tilde{D}(z) = I_m, \quad C(1) = D(1) = I_m.$$

Proof. It is straightforward to verify that a pair of (L, R) with the above form of factorization is indeed a biorthogonal wavelet matrix pair.

For the converse part, the proof comes from Lemma 3.1, Theorem 3.2, and Theorem 3.3. □

Remark 2. Note that in the above theorem, d can be 0, i.e., there are no $V_i(z)$ factors. Such a biorthogonal wavelet matrix pair is given in (3.7).

For two biorthogonal wavelet matrix pairs of rank m, (L^1, R^1) and (L^2, R^2), that have the same characteristic Haar matrix pair, we define the *Pollen product* by the formula

$$L^1 \diamond_{H_L} L^2 := L, \quad R^1 \diamond_{H_R} R^2 := R$$

if their Laurent series satisfy

$$L(z) = L^1(z) H_L^{-1} L^2(z), \quad R(z) = R^1(z) H_R^{-1} R^2(z),$$

where

$$H_L = \chi(L^1) = \chi(L^2), \quad H_R = \chi(R^1) = \chi(R^2).$$

The characteristic Haar matrix pair of (L, R) is also (H_L, H_R). The set of biorthogonal wavelet matrix pairs of the same characteristic Haar matrix pair is a group under this product.

THEOREM 3.5. *Given a biorthogonal Haar wavelet matrix pair (H_L, H_R), let* $\mathrm{WM}(H_L, H_R)$ *be the collection of biorthogonal wavelet matrix pairs whose characteristic Haar matrix pair is (H_L, H_R). Then* $\mathrm{WM}(H_L, H_R)$ *is a group with the noncommutative product*

$$((L^1, R^1), (L^2, R^2)) \longmapsto \left(L^1 \diamond_{H_L} L^2, R^1 \diamond_{H_R} R^2\right)$$

and the unit element is (H_L, H_R).

The proof of this theorem is elementary and is omitted.

With the Pollen product, we can rephrase Theorem 3.4 into a factorization inside the group $\mathrm{WM}(H_L, H_R)$.

THEOREM 3.6. *For a biorthogonal wavelet matrix pair (L, R) of rank m and genus g, there exist biorthogonal wavelet matrix pairs $(L^1, R^1), \ldots, (L^d, R^d)$, and (C^1, D^1) such that*

$$L = z^{-k_0} L^1 \diamond_{H_L} L^2 \diamond_{H_L} \cdots \diamond_{H_L} L^d \diamond_{H_L} C^1, \quad R = z^{-k_0} R^1 \diamond_{H_R} R^2 \diamond_{H_R} \cdots \diamond_{H_R} R^d \diamond_{H_R} D^1,$$

where $k_0 \in \mathbf{Z}$, $H_L = \chi(L), H_R = \chi(R)$, and

$$L^i(z) = V_i(z) H_L = (I_m - v_i {v_i}^* + v_i {v_i}^* z^{-1}) H_L,$$

$$R^i(z) = V_i(z) H_R = (I_m - v_i {v_i}^* + v_i {v_i}^* z^{-1}) H_R,$$

with unit column vectors $v_1, \ldots, v_d$, $i = 1, \ldots, d$, and

$$C^1(z) = \sum_{k=0}^{k_c} C_k^1 z^{-k}, \quad D^1(z) = \sum_{k=k_d}^{0} D_k^1 z^{-k},$$

with $k_c \leq g - 1$, and

$$C^1(z) \tilde{D^1}(z) = m I_m.$$

4. The pseudo identity matrix pair. In the factorization process in the previous section, a biorthogonal wavelet matrix pair is reduced to a matrix pair $(C(z), D(z))$ with the following four properties:

1. $C(z) = \sum_{k=0}^{k_c} C_k z^{-k}, D(z) = \sum_{k=k_d}^{0} D_k z^{-k}$,
2. $C(z)\tilde{D}(z) = I_m$,
3. $\det(C(z)) = \det(D(z)) = 1$,
4. $C(1) = D(1) = I_m$.

The above four properties are redundant. With a $C(z)$ satisfying properties 1 and 3, there always exists a $D(z)$ satisfying properties 1, 2, and 3. With property 2, all other three properties of $D(z)$ can be derived from those of $C(z)$. With properties 1 and 2, property 3 can be derived from property 4. Actually, properties 3 and 4 are just normalization conditions. In the rest of this section, we will concentrate on properties 1 and 2.

In an orthogonal setting (i.e., $C = D$), one can prove that $C(z)$ and $D(z)$ must be equal to the identity matrix I_m. However, in a biorthogonal setting, there exist nontrivial pairs of $(C(z), D(z))$ satisfying all four properties. Here is an example:

$$C(z) = \begin{pmatrix} -1 & 1 \\ -4 & 3 \end{pmatrix} + \begin{pmatrix} 2 & -1 \\ 4 & -2 \end{pmatrix} z^{-1}, \quad D(z) = \begin{pmatrix} -2 & -4 \\ 1 & 2 \end{pmatrix} z + \begin{pmatrix} 3 & 4 \\ -1 & -1 \end{pmatrix}.$$

For convenience, we will call $(C(z), D(z))$ a pseudo identity matrix pair.

Definition 4.1. *A matrix pair* $(C(z), D(z))$ *is a* pseudo identity matrix pair *if*

$$C(z) = \sum_{k=0}^{k_c} C_k z^{-k}, \quad D(z) = \sum_{k=k_d}^{0} D_k z^{-k}, \quad C(z)\tilde{D}(z) = I_m, \quad C(1) = D(1) = I_m.$$

In this section, we will study the structure of the pseudo identity matrix pair. When the size is of $m \times 2m$ (which is the minimal length for a nontrivial pair), we have the following lemma.

Lemma 4.2. *If two* $m \times 2m$ *matrices* C *and* D *satisfy*

$$C(z) = C_0 + C_1 z^{-1}, \quad D(z) = D_{-1} z + D_0, \quad C(z)\tilde{D}(z) = I_m, \quad C(1) = D(1) = I_m,$$

then there exists an $m \times m$ *nilpotent matrix* N, $N^2 = 0$, *such that*

$$C(z) = I_m - N + N z^{-1}, \quad D(z) = -N^* z + I_m + N^*.$$

Proof. Since $C(1) = D(1) = I_m$, we have

$$C_0 = I_m - C_1, \quad D_0 = I_m - D_{-1}.$$

Now

$$\begin{aligned} C(z)\tilde{D}(z) &= \left(I_m - C_1 + C_1 z^{-1}\right)\left(I_m - D_{-1}^* + D_{-1}^* z^{-1}\right) \\ &= I_m - C_1 - D_{-1}^* + C_1 D_{-1}^* + \left(D_{-1}^* + C_1 - 2C_1 D_{-1}^*\right) z^{-1} + C_1 D_{-1}^* z^{-2}. \end{aligned}$$

It follows that

$$C_1 D_{-1}^* = 0, \tag{4.1}$$

$$D_{-1}^* + C_1 - 2C_1 D_{-1}^* = 0. \tag{4.2}$$

Substituting (4.1) into (4.2), we have

$$D^*_{-1} = -C_1.$$

Combined with (4.1), it follows that

$$(C_1)^2 = 0. \qquad \square$$

For an $m \times m$ nilpotent matrix N, with $N^2 = 0$, define the Laurent matrix pair

$$L_N(z) := I_m - N + Nz^{-k}, \quad R_N(z) := -N^* z^k + I_m + N^*, \tag{4.3}$$

and let (L_N, R_N) be the corresponding matrix pair. We will say that a matrix pair (L_N, R_N) of form (4.3) is a *primitive pseudo identity matrix pair of degree k*. It's easy to verify that (L_N, R_N) is indeed a pseudo identity matrix pair. In addition, for any positive integer n,

$$(L_N(z))^n = I_m - nN + nNz^{-k}, \quad (R_N(z))^n = -nN^* z^k + I_m + nN^*.$$

In the case of $m = 2$, $N^2 = 0$ implies

$$N = \begin{pmatrix} w & u \\ v & -w \end{pmatrix}$$

for some w, u, v satisfying $uv = -w^2$, which means N has parameter freedom of 2. So when $m = 2$, a primitive pseudo identity matrix pair will be of the form

$$L_N(z) = \begin{pmatrix} 1 - w(1 - z^{-k}) & -u(1 - z^{-k}) \\ -v(1 - z^{-k}) & 1 + w(1 - z^{-k}) \end{pmatrix},$$

$$R_N(z) = \begin{pmatrix} 1 + w(-z^k + 1) & v(-z^k + 1) \\ u(-z^k + 1) & 1 - w(-z^k + 1) \end{pmatrix},$$

with $uv = -w^2$.

Lemma 4.3. *For*

$$C(z) = \sum_{k=0}^{k_c} C_k z^{-k}, \quad D(z) = \sum_{k=k_d}^{0} D_k z^{-k},$$

if

$$C(z)\tilde{D}(z) = I_m$$

and both C_{k_c-1} and D_{k_d+1} are invertible matrices, then there exists a nilpotent matrix N, with $N^2 = 0$, such that

$$C(z) = \left(\sum_{k=0}^{k_c-1} C^1_k z^{-k}\right) L_N(z), \quad D(z) = \left(\sum_{k=k_d+1}^{0} D^1_k z^{-k}\right) R_N(z),$$

with

$$L_N(z) = I_m - N + Nz^{-1}, \quad R_N(z) = -N^* z + I_m + N^*.$$

Proof. Since

$$C(z)\tilde{D}(z) = \left(\sum_{k=0}^{k_c} C_k z^{-k}\right)\left(\sum_{k=k_d}^{0} D_k^* z^k\right) = \left(\sum_{k=0}^{k_c} C_k z^{-k}\right)\left(\sum_{k=0}^{-k_d} D_{-k}^* z^{-k}\right) = I_m,$$

we have

$$C_{k_c} D_{k_d}^* = 0, \quad C_{k_c} D_{k_d+1}^* + C_{k_c-1} D_{k_d}^* = 0.$$

Now set

$$N = (C_{k_c-1})^{-1} C_{k_c} = -D_{k_d}^* (D_{k_d+1}^*)^{-1},$$

then

$$N^2 = 0, \quad C_{k_c} N = 0, \quad N D_{k_d}^* = 0,$$

$$-C_{k_c-1} N + C_{k_c} = 0, \quad N D_{k_d+1}^* + D_{k_d}^* = 0.$$

So

$$C(z)(L_N(z))^{-1} = \left(\sum_{k=0}^{k_c} C_k z^{-k}\right)(I_m + N - N z^{-1}) = \sum_{k=0}^{k_c-1} C_k^1 z^{-k},$$

$$D(z)(R_N(z))^{-1} = \left(\sum_{k=k_d}^{0} D_k z^{-k}\right)(N^* z + I_m - N^*) = \sum_{k=k_d+1}^{0} D_k^1 z^{-k}.$$

That is,

$$C(z) = C^1(z) L_N(z) = \left(\sum_{k=0}^{k_c-1} C_k^1 z^{-k}\right) L_N(z),$$

$$D(z) = D^1(z) R_N(z) = \left(\sum_{k=k_d+1}^{0} D_k^1 z^{-k}\right) R_N(z). \qquad \square$$

Notice that the genus of C^1 and D^1 is reduced by one (from $k_c + 1$ to k_c and from $-k_d + 1$ to $-k_d$) simultaneously, compared with C and D. The reduction from (C, D) to (C^1, D^1) totally depends on whether or not C_{k_c-1} and D_{k_d+1} are invertible. We formulate this condition as follows.

Condition 1. *For $m = 2$,*

$$C(z) = \sum_{k=0}^{k_c} C_k z^{-k}, \quad D(z) = \sum_{k=k_d}^{0} D_k z^{-k},$$

if

$$C(z)\tilde{D}(z) = I_2,$$

and p is the smallest positive integer such that $C_{k_c-p} \neq 0$, then both C_{k_c-p} and D_{k_d+p} are invertible matrices.

Because of the symmetry of C and D, if p' is the smallest positive integer such that $D_{k_d+p'} \neq 0$, then the above condition asserts that both $C_{k_c-p'}$ and $D_{k_d+p'}$ are invertible.

THEOREM 4.4. *If Condition 1 is true, then for any pair (C, D) of rank 2 satisfying*

$$C(z) = \sum_{k=0}^{k_c} C_k z^{-k}, \quad D(z) = \sum_{k=k_d}^{0} D_k z^{-k}, \quad C(z)\tilde{D}(z) = I_2, \quad C(1) = D(1) = I_2,$$

there exist nilpotent matrices $N_1, N_2, \ldots, N_q, N_i^2 = 0, i = 1, 2, \ldots, q$, such that

$$C(z) = L_{N_q}(z) \cdots L_{N_2}(z) L_{N_1}(z), \quad D(z) = R_{N_q}(z) \cdots R_{N_2}(z) R_{N_1}(z),$$

with

$$L_{N_i}(z) = I_2 - N_i + N_i z^{-k_i}, \quad R_{N_i}(z) = -N_i^* z^{k_i} + I_2 + N_i^*, \quad \text{for } i = 1, 2, \ldots, q,$$

and

$$k_1 + k_2 + \cdots + k_q = k_c = -k_d.$$

Proof. Assume k_1 is the smallest integer such that either $C_{k_c-k_1}$ or $D_{k_d+k_1}$ is nonzero. By Condition 1, both $C_{k_c-k_1}$ and $D_{k_d+k_1}$ are invertible matrices. Similar to the proof of Lemma 4.3, from

$$\begin{aligned} C(z)\tilde{D}(z) &= \left(\sum_{k=0}^{k_c} C_k z^{-k}\right)\left(\sum_{k=k_d}^{0} D_k^* z^k\right) \\ &= \left(\sum_{k=0}^{k_c-k_1} C_k z^{-k} + C_{k_c} z^{-k_c}\right)\left(\sum_{k=0}^{-k_d-k_1} D_{-k}^* z^{-k} + D_{k_d}^* z^{-k_d}\right) = I_2, \end{aligned}$$

we have

$$C_{k_c} D_{k_d}^* = 0, \quad C_{k_c} D_{k_d+k_1}^* + C_{k_c-k_1} D_{k_d}^* = 0.$$

Define

$$N_1 = (C_{k_c-k_1})^{-1} C_{k_c} = -D_{k_d}^* (D_{k_d+k_1}^*)^{-1},$$

$$L_{N_1}(z) = I_2 - N_1 + N_1 z^{-k_1}, \quad R_{N_1}(z) = -N_1^* z^{k_1} + I_2 + N_1^*.$$

Then it can be easily verified that

$$N_1^2 = 0,$$

$$C(z) = \left(\sum_{k=0}^{k_c-k_1} C_k^1 z^{-k}\right) L_{N_1}(z) = C^1(z) L_{N_1}(z),$$

$$D(z) = \left(\sum_{k=k_d+k_1}^{0} D_k^1 z^{-k} \right) R_{N_1}(z) = D^1(z) R_{N_1}(z).$$

The genus of C^1 and D^1 is reduced by k_1, compared with C and D. Now repeating the genus reduction procedure on the new pair (C^1, D^1), one can eventually obtain

$$C(z) = L_{N_q}(z) \cdots L_{N_2}(z) L_{N_1}(z), \quad D(z) = R_{N_q}(z) \cdots R_{N_2}(z) R_{N_1}(z),$$

with $k_1 + k_2 + \cdots + k_q = k_c = -k_d$. □

Remark 3. Note that Condition 1 is actually a very strong condition to guarantee the existence of a nilpotent matrix for the genus reduction procedure. All we need for such a genus reduction is a weaker condition

$$-C_{k_c-p} N + C_{k_c} = 0, \quad N D^*_{k_d+p} + D^*_{k_d} = 0$$

for some nilpotent matrix N, $N^2 = 0$. The advantage of Condition 1 is that it's much easier to check.

THEOREM 4.5. *If* Condition 1 *is true, and a pair of matrices* (L, R) *is a biorthogonal wavelet matrix pair of rank* 2*, then there exist primitive paraunitary matrices* $V_1, \ldots, V_d$*,* $d \geq 0$*, and primitive pseudo identity matrix pairs* (L_{N_1}, R_{N_1})*,* (L_{N_2}, R_{N_2})*,* $\ldots$*,* (L_{N_q}, R_{N_q})*,* $q \geq 0$*, such that*

$$L(z) = z^{-k_0} V_1(z) \cdots V_d(z) L_{N_q}(z) \cdots L_{N_1}(z) \begin{pmatrix} 1 & 0 \\ 0 & c \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix},$$

$$R(z) = z^{-k_0} V_1(z) \cdots V_d(z) R_{N_q}(z) \cdots R_{N_1}(z) \begin{pmatrix} 1 & 0 \\ 0 & c^{-1} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix},$$

where $k_0 \in \mathbf{Z}$*,* $d = b - 2k_0$*,* b *is the exponent of* $\det(L(z))$*,* $c \neq 0$ *is a constant.*

5. Conversion between orthogonal and biorthogonal wavelets. Theorem 3.4 tells us that a biorthogonal wavelet matrix pair (L, R) can be decomposed into four components:

$$L(z) = V(z) C(z) \begin{pmatrix} 1 & 0 \\ 0 & G \end{pmatrix} \mathbf{H}, \quad R(z) = V(z) D(z) \begin{pmatrix} 1 & 0 \\ 0 & (G^{-1})^* \end{pmatrix} \mathbf{H},$$

with

$$V(z) = z^{-k_0} V_1(z) V_2(z) \cdots V_d(z),$$

$(C(z), D(z))$ is the pseudo identity matrix pair discussed in section 4, G is an invertible matrix, and $\mathbf{H}$ is the canonical Haar matrix. Only the first and second components $V(z), (C(z), D(z))$ contain the variable z. The fourth component $\mathbf{H}$ is a constant matrix, depending only on the rank m. The third component G doesn't contain z, and it can be easily determined by the value of $L(1)$ (or equivalently $R(1)$), since

$$L(1) = \begin{pmatrix} 1 & 0 \\ 0 & G \end{pmatrix} \mathbf{H}.$$

When looking at the determinant, if

$$\det(L(z)) = c_l z^{-b}, \quad \det(R(z)) = c_r z^{-b}, \quad c_l c_r = m,$$

then

$$\det(V(z)) = z^{-b}, \quad \det(C(z)) = \det(D(z)) = 1, \quad \det(G) = c_l/\sqrt{m}, \quad \det(\mathbf{H}) = \sqrt{m}.$$

Thus these four components have distinctive determinant values from each other.

Recall that in the theory of orthogonal wavelet matrices [13, 16, 25, 26, 28, 29], we have the following characterization result.

THEOREM 5.1 (orthogonal factorization theorem). *An $m \times mg$ matrix A is an orthogonal wavelet matrix if and only if there exist primitive paraunitary matrices $V_1, \ldots, V_d, d \geq 0$, such that*

$$A(z) = z^{-k} V_1(z) V_2(z) \cdots V_d(z) \begin{pmatrix} 1 & 0 \\ 0 & U \end{pmatrix} \mathbf{H},$$

where $k \in \mathbf{Z}$, $U \in U_{m-1}$ is a unitary matrix, that is, $U^* U = I_{m-1}$, and $\mathbf{H}$ is the canonical Haar matrix of rank m.

Comparing Theorem 5.1 with Theorem 3.4, it is now clear that the difference between a biorthogonal wavelet matrix pair and an orthogonal wavelet matrix lies in the pseudo identity matrix pair $(C(z), D(z))$ and the invertible matrix G. If $C(z)$ and $D(z)$ are both equal to the identity matrix and G is unitary, then a biorthogonal wavelet matrix pair (L, R) is actually an orthogonal wavelet matrix (and $L = R$). Thus to "orthogonalize" a biorthogonal wavelet matrix pair (L, R), one needs to and only needs to throw away the pseudo identity matrix pair $(C(z), D(z))$ and "orthogonalize" the invertible matrix G.

In linear algebra, a standard method to convert an invertible matrix to a unitary matrix is the Schmidt orthogonalization method. Given an invertible matrix $G \in GL_{m-1}$, its $m-1$ row vectors $\alpha_1, \ldots, \alpha_{m-1}$ are linearly independent. Now define

$$e_1 := \frac{\alpha_1}{||\alpha_1||};$$

then e_1 is a unit vector and has the same direction as α_1. From α_2 we construct a unit vector e_2 perpendicular to e_1 by

$$\beta_2 = \alpha_2 - \langle \alpha_2, e_1 \rangle e_1, \quad e_2 := \frac{\beta_2}{||\beta_2||}.$$

Proceeding in this fashion, we can define $e_3, \ldots, e_{m-1}$ by

$$\beta_3 = \alpha_3 - \langle \alpha_3, e_1 \rangle e_1 - \langle \alpha_3, e_2 \rangle e_2, \quad e_3 := \frac{\beta_3}{||\beta_3||},$$

$$\vdots$$

$$\beta_{m-1} = \alpha_{m-1} - \langle \alpha_{m-1}, e_1 \rangle e_1 - \cdots - \langle \alpha_{m-1}, e_{m-2} \rangle e_{m-2}, \quad e_{m-1} := \frac{\beta_{m-1}}{||\beta_{m-1}||}.$$

Then the matrix

$$U := \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{m-1} \end{pmatrix}$$

is a unitary matrix.

Thus using the Schmidt orthogonalization method, we can convert a biorthogonal wavelet matrix pair to an orthogonal matrix with the same rank. (Note that orthogonalizing the analysis matrix and synthesis matrix will give the same orthogonal wavelet matrix.) Conversely we can retrieve invertible matrices from a unitary matrix by setting up a mapping from unitary matrices to invertible matrices. By adding an additional pseudo identity matrix pair, one can construct biorthogonal wavelet matrix pairs from orthogonal wavelet matrices.

In general, any mapping from the invertible matrix group to the unitary matrix group will define a mapping from biorthogonal wavelet matrix pairs to orthogonal wavelet matrices,

$$\begin{aligned}(L, R) &\longrightarrow A,\\ (C(z), D(z)) &\longmapsto I_m,\\ G &\longmapsto U.\end{aligned}$$

Next, we will discuss one interesting case, namely, how to preserve the vanishing moments. It can be proved that the vanishing moment conditions of scaling functions and/or wavelet functions can be translated into equivalent conditions on the wavelet matrix (or matrix pair). First, let's assume the orthogonal wavelet matrix A has vanishing moments up to order n on the scaling function $\phi(x)$, that is,

$$\sum_j a_{0,j} = m, \tag{5.1}$$

$$\sum_j a_{0,j} j^k = 0, \quad k = 1, \ldots, n. \tag{5.2}$$

The zeroth order vanishing moment condition (5.1) is already in the definition of an orthogonal wavelet matrix. For the first order vanishing moment ($k = 1$), since

$$\begin{aligned}\sum_j a_{0,j} j &= \sum_j \left(\sum_{l=0}^{m-1} a_{0,mj+l}(mj + l) \right)\\ &= m \sum_j j \left(\sum_{l=0}^{m-1} a_{0,mj+l} \right) + \sum_{l=1}^{m-1} l \left(\sum_j a_{0,mj+l} \right)\end{aligned}$$

and

$$A'(z) = \left(\sum_j A_j z^{-j} \right)' = -\sum_j j A_j z^{-j-1},$$

$$A'(1) = -\sum_j j A_j,$$

it follows that the first order vanishing moment of the scaling function $\phi(x)$ is equivalent to a linear equation on the entries of the first row of $A(1)$ and $A'(1)$. Similarly, one can convert all vanishing moment conditions (up to order n) of the scaling function and/or wavelet functions into some linear equations on the entries of

$A(1), A'(1), A^{(2)}(1), \ldots, A^{(n)}(1)$. Equivalently, with a biorthogonal wavelet matrix pair (L, R), one can also convert all vanishing moment conditions into linear equations of $L(1), L'(1), L^{(2)}(1), \ldots,$ $L^{(n)}(1)$, and $R(1), R'(1), R^{(2)}(1), \ldots, R^{(n)}(1)$.

Thus a sufficient condition to preserve the vanishing moment conditions when constructing a biorthogonal wavelet matrix pair (L, R) from an orthogonal wavelet matrix A is that

$$L(1) = R(1) = A(1),$$

$$L'(1) = R'(1) = A'(1),$$

$$\vdots$$

$$L^{(n)}(1) = R^{(n)}(1) = A^{(n)}(1).$$

One could choose a $(C(z), D(z))$ component pair such that

$$C(1) = D(1) = I_m,$$

$$C^{(k)}(1) = D^{(k)}(1) = 0, \quad k = 1, \ldots, n.$$

Then the resulting biorthogonal wavelet matrix pair (L, R)

$$L(z) = A(z)C(z), \quad R(z) = A(z)D(z)$$

will preserve all vanishing moments from A. For example, one can take

$$C(z) = I_m + N(1 - z^{-1})^{n+1}, \quad D(z) = I_m - N^*(1 - z)^{n+1},$$

with an $m \times m$ nilpotent matrix N, satisfying $N^2 = 0$.

6. Examples. In this section we will give a simple example to illustrate the factorization of biorthogonal wavelet matrices. We take the 3-5 biorthogonal wavelet matrix pair from [8]. The analysis matrix L is given by

$$L = \begin{pmatrix} -\frac{1}{4} & \frac{1}{2} & \frac{3}{2} & \frac{1}{2} & -\frac{1}{4} & 0 \\ \frac{1}{2} & -1 & \frac{1}{2} & 0 & 0 & 0 \end{pmatrix}.$$

Its Laurent series is

$$L(z) = \begin{pmatrix} -\frac{1}{4} & \frac{1}{2} \\ \frac{1}{2} & -1 \end{pmatrix} + \begin{pmatrix} \frac{3}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix} z^{-1} + \begin{pmatrix} -\frac{1}{4} & 0 \\ 0 & 0 \end{pmatrix} z^{-2}.$$

The four component decomposition will be

$$L(z) = V_1(z)C(z) \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \mathbf{H},$$

where

$$V_1(z) = I_2 - v_1 v_1^t + v_1 v_1^t z^{-1} = \begin{pmatrix} \frac{1}{5} & -\frac{2}{5} \\ -\frac{2}{5} & \frac{4}{5} \end{pmatrix} + \begin{pmatrix} \frac{4}{5} & \frac{2}{5} \\ \frac{2}{5} & \frac{1}{5} \end{pmatrix} z^{-1},$$

with

$$v_1 = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{pmatrix},$$

and

$$C(z) = \begin{pmatrix} \frac{41}{40} & \frac{1}{8} \\ \frac{1}{5} & 1 \end{pmatrix} + \begin{pmatrix} 0 & -\frac{1}{10} \\ -\frac{1}{4} & -\frac{1}{20} \end{pmatrix} z^{-1} + \begin{pmatrix} -\frac{1}{40} & -\frac{1}{40} \\ \frac{1}{20} & \frac{1}{20} \end{pmatrix} z^{-2}, \quad \mathbf{H} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}.$$

Consequently, the synthesis matrix R can be factored as

$$R(z) = V_1(z)D(z) \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \mathbf{H},$$

where $(C(z), D(z))$ is a pseudo identity matrix pair. Now since C_1 in $C(z)$ is invertible, based on the techniques in Lemma 4.3, $C(z)$ can be further factored as

$$\begin{aligned} C(z) &= L_{N_1}(z) \cdot L_{N_2}(z) \\ &= \left(I_2 - \begin{pmatrix} \frac{1}{5} & \frac{1}{10} \\ -\frac{2}{5} & -\frac{1}{5} \end{pmatrix} + \begin{pmatrix} \frac{1}{5} & \frac{1}{10} \\ -\frac{2}{5} & -\frac{1}{5} \end{pmatrix} z^{-1} \right) \\ &\quad \cdot \left(I_2 - \begin{pmatrix} -\frac{1}{4} & -\frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix} + \begin{pmatrix} -\frac{1}{4} & -\frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix} z^{-1} \right). \end{aligned}$$

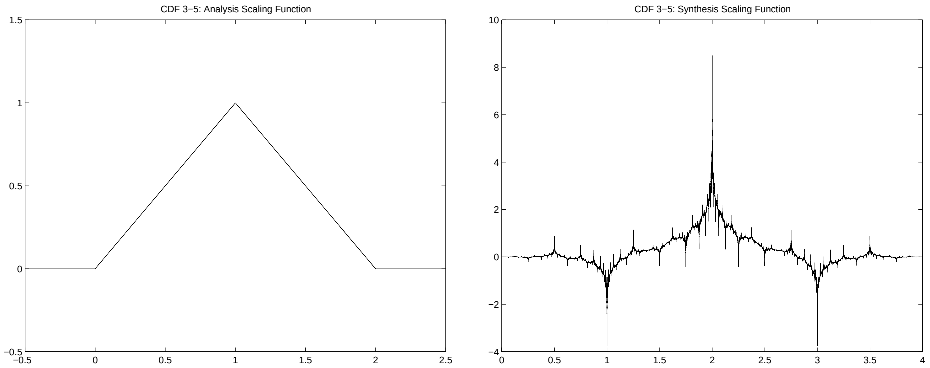
Thus

$$\begin{aligned} D(z) = &\left(- \begin{pmatrix} \frac{1}{5} & -\frac{2}{5} \\ \frac{1}{10} & -\frac{1}{5} \end{pmatrix} z + I_2 + \begin{pmatrix} \frac{1}{5} & -\frac{2}{5} \\ \frac{1}{10} & -\frac{1}{5} \end{pmatrix} \right) \\ &\cdot \left(- \begin{pmatrix} -\frac{1}{4} & \frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} \end{pmatrix} z + I_2 + \begin{pmatrix} -\frac{1}{4} & \frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} \end{pmatrix} \right). \end{aligned}$$

To derive an orthogonal wavelet matrix A from L, one can simply throw away $C(z)$ and get

$$A(z) = V_1(z) \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} -\frac{1}{5} & \frac{3}{5} \\ \frac{2}{5} & -\frac{6}{5} \end{pmatrix} + \begin{pmatrix} \frac{6}{5} & \frac{2}{5} \\ \frac{3}{5} & \frac{1}{5} \end{pmatrix} z^{-1}.$$

For illustration, we include the graphs of the analysis scaling function and the synthesis scaling function constructed from the 3-5 biorthogonal wavelet matrix pair (L, R), and the graph of the scaling function constructed from the orthogonal wavelet matrix A in Figures 1 and 2, respectively. The analysis scaling function in Figure 1 is a linear spline function, and it is in $C^{1-\epsilon}$, for any $0 < \epsilon < 1$, while the synthesis scaling function is not even continuous. The scaling function of the "orthogonalized" wavelet matrix A in Figure 2 is in $C^{0.3219}$. Thus in this example, the smoothness of the "orthogonalized" scaling function is between those of the analysis scaling function and the synthesis scaling function of the original biorthogonal wavelet.



(a) Analysis Scaling Function (b) Synthesis Scaling Function

FIG. 1. *Scaling functions of the* 3-5 *biorthogonal pair.*

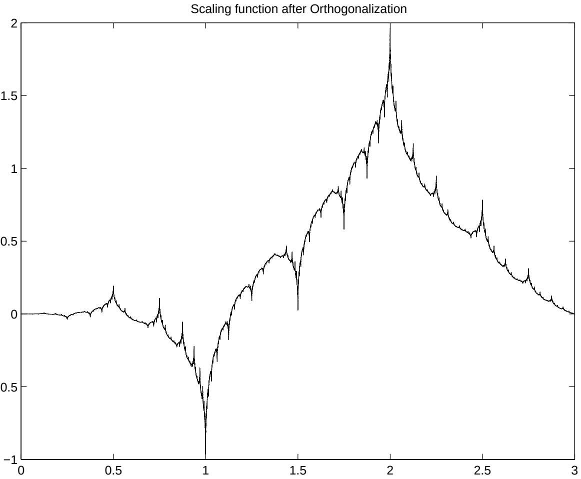


FIG. 2. *Scaling function of the "orthogonalized" wavelet matrix.*

7. Conclusions. In this paper we study the algebraic and geometric structure of the space of compactly supported biorthogonal wavelets. The wavelet matrix can be real-valued, complex-valued, or in any subfield of complex numbers closed under the complex conjugation. We present a complete characterization of biorthogonal wavelet matrix pairs. The conversion between orthogonal and biorthogonal wavelets is provided. We also discuss how to preserve the vanishing moment condition in such a conversion. There are still several open problems in the factorization of biorthogonal wavelet matrix pairs, such as how to construct symmetric biorthogonal wavelets from the factorization formula, the smoothness estimate of biorthogonal wavelets from the factorization formula, etc. We are currently investigating these problems and the progress will be reported in a forthcoming paper.

Acknowledgments. We would like to thank Clayton Scott of Rice University for valuable assistance. We also thank the reviewers. Their comments have enhanced the paper considerably. Tian and Wells would like to thank their colleagues in the Computational Mathematics Laboratory of Rice University, from where they are on

leave of absence.

After completion of this work, Fritz Keinert [17] showed to us a counterexample of Condition 1 (and even the weaker condition) as follows:

$$\begin{aligned} C(z) &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} (z^{-1} - z^{-2}), \\ D(z) &= \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} (z^2 - z) + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \end{aligned}$$

For the above pseudo identity matrix pair (C, D), they can be factored as

$$\begin{aligned} C(z) &= \left(I_2 - N + Nz^{-1}\right) \cdot \left(I_2 + N - Nz^{-2}\right), \\ D(z) &= \left(N^* z^2 + I_2 - N^*\right) \cdot \left(-N^* z + I_2 + N^*\right), \end{aligned}$$

where

$$N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Thus they could still be factored as the product of primitive pseudo identity matrix pairs even though Condition 1 could not hold. Finding the necessary and sufficient condition such that a pseudo identity matrix pair could be factored as the product of primitive pseudo identity matrix pairs remains an open problem. Keinert develops a numerical algorithm to generate all pseudo identity matrix pairs by an explicit parametrization. Interested readers are referred to [18]. He also drew our attention to [15], which obtained a similar result to Theorem 3.6 in this paper. We would like to thank Keinert for his helpful comments and for sharing his work with us.

REFERENCES

[1] G. Battle, *A block spin construction of ondelettes. Part* I*: Lemarié functions*, Comm. Math. Phys., 110 (1987), pp. 610–615.

[2] S. Borac and R. Seiler, *Loop Group Factorization of Biorthogonal Wavelet Bases*, preprint.

[3] C. S. Burrus, R. A. Gopinath, and H. Guo, *Introduction to Wavelets and Wavelet Transforms*, Prentice-Hall, Englewood Cliffs, NJ, 1997.

[4] C. K. Chui, *An Introduction to Wavelets*, Academic Press, Boston, MA, 1992.

[5] C. K. Chui, *On cardinal spline wavelets*, in Wavelets and Their Applications, Jones and Bartlett, Boston, 1992, pp. 419–438.

[6] C. K. Chui and J. Z. Wang, *A general framework of compactly supported splines and wavelets*, J. Approx. Theory, 71 (1992), pp. 263–304.

[7] A. Cohen, *Ondelettes, analyses multirésolutions et filtres miroir en quadrature*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 439–459.

[8] A. Cohen, I. Daubechies, and J.-C. Feauveau, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 45 (1992), pp. 485–560.

[9] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 906–966.

[10] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA, 1992.

[11] I. Daubechies and W. Sweldens, *Factoring wavelet transforms into lifting steps*, J. Fourier Anal. Appl., 4 (1998), pp. 245–267.

[12] A. Grossmann and J. Morlet, *Decomposition of Hardy functions into square integrable wavelets of constant shape*, SIAM J. Math. Anal., 15 (1984), pp. 723–736.

[13] P. N. Heller, H. L. Resnikoff, and R. O. Wells, Jr., *Wavelet matrices and the representation of discrete functions*, in Wavelets. A Tutorial in Theory and Applications, C. K. Chui, ed., Academic Press, Boston, 1992, pp. 15–50.

[14] M. Holschneider, *Wavelets. An Analysis Tool*, Oxford University Press, New York, 1995.

[15] J. Kautsky and R. Turcajová, *Discrete biorthogonal wavelet transforms as block circulant matrices*, Linear Algebra Appl., 223/224 (1995), pp. 393–413.

[16] J. Kautsky and R. Turcajová, *Pollen product factorization and construction of higher multiplicity wavelets*, Linear Algebra Appl., 222 (1995), pp. 241–260.

[17] F. Keinert, *private communication*, 2000.

[18] F. Keinert, *Parametrization of unimodular matrix polynomials*, Linear Algebra Appl., submitted.

[19] A. Klappenecker, M. Holschneider, and K. Flornes, *Two-channel perfect reconstruction filter banks over commutative rings*, Appl. Comput. Harmon. Anal., 8 (2000), pp. 113–121.

[20] W. M. Lawton, *Necessary and sufficient conditions for constructing orthogonal wavelet bases*, J. Math. Phys., 32 (1991), pp. 57–61.

[21] P. G. Lemarié, *Une nouvelle bade d'ondelettes de* $L^2(\mathbf{R}^n)$, Math. Pure et Appl., 67 (1988), pp. 227–236.

[22] S. G. Mallat, *Multiresolution approximation and wavelet orthonormal bases of* $L^2(\mathbf{R})$, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87.

[23] Y. Meyer, *Wavelets and Operators*, Cambridge University Press, Cambridge, UK, 1992.

[24] R. E. A. C. Paley and N. Wiener, *Fourier Transforms in the Complex Domain*, American Mathematical Society, Providence, RI, 1934.

[25] D. Pollen, $SU_I(2, F[z, 1/z])$ *for* F *a subfield of* $\mathbf{C}$, J. Amer. Math. Soc., 3 (1990), pp. 611–624.

[26] H. L. Resnikoff and R. O. Wells, Jr., *Wavelet Analysis and the Scalable Structure of Information*, Springer-Verlag, New York, 1998.

[27] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambridge Press, Wellesley, MA, 1995.

[28] J. Tian and R. O. Wells, Jr., *An algebraic structure of orthogonal wavelet space*, Appl. Comput. Harmon. Anal., 8 (2000), pp. 223–248.

[29] P. P. Vaidyanathan, T. Q. Nguyen, Z. Doğanata, and T. Saramäki, *Improved technique for design of perfect reconstruction FIR QMF banks with lossless polyphase matrices*, IEEE Trans. ASSP, 37 (1989), pp. 1042–1056.

[30] M. Vetterli and C. Herley, *Wavelets and filter banks: Theory and design*, IEEE Trans. ASSP, 40 (1992), pp. 2207–2232.

[31] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[32] M. V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*, A. K. Peters, Ltd., Wellesley, MA, 1993.

SIAM J. MATH. ANAL.
Vol. 33, No. 1, pp. 216–239

ASYMPTOTIC BEHAVIOR OF SOLUTIONS TO THE SYSTEM OF COMPRESSIBLE ADIABATIC FLOW THROUGH POROUS MEDIA[*]

KENJI NISHIHARA[†] AND MASATAKA NISHIKAWA[‡]

Abstract. Hsiao and Serre in [*Chinese Ann. Math. Ser. B,* 16B (1995), pp. 1–14] showed the solution to the system

$$\begin{cases} v_t - u_x = 0, & (t,x) \in R_+ \times R, \\ u_t + p(v,s)_x = -\alpha u, & \alpha > 0, \\ s_t = 0 \end{cases}$$

with initial data

$$(v,u,s)(0,x) = (v_0,u_0,s_0)(x) \to (\underline{v}, u_\pm, \underline{s}) \qquad \text{as} \quad x \to \pm\infty$$

tends to the following nonlinear parabolic equation time-asymptotically:

$$\begin{cases} \tilde{v}_t = -\frac{1}{\alpha} p(\tilde{v}, s_0)_{xx}, & (t,x) \in R_+ \times R, \\ \tilde{u} = -\frac{1}{\alpha} p(\tilde{v}, s_0)_x. \end{cases}$$

In this paper we find its convergence rate, which will be optimal.

Key words. asymptotic behavior, the system of compressible adiabatic flow, convergence rate

AMS subject classifications. 35L65, 35L67, 76L05

PII. S003614109936467X

1. Introduction.

We consider the Cauchy problem for the equation of the form

$$\begin{cases} v_t - u_x = 0, & (t,x) \in R_+ \times R, \\ u_t + p(v,s)_x = -\alpha u, & \alpha > 0, \\ s_t = 0, \end{cases} \tag{1.1}$$

which can be used to model the adiabatic gas flow through porous media. Here v is the specific volume, u denotes the velocity, s stands for the entropy, and p denotes the pressure with $p > 0$, $p_v < 0$ for $v > 0$. A typical example of p is $p(v,s) = (\gamma - 1)v^{-\gamma}e^s$ $(\gamma > 1)$.

In Hsiao and Serre [2, 3], it has been proved that the solution of the Cauchy problem (1.1) with

$$(v,u,s)(0,x) = (v_0,u_0,s_0)(x) \to (v_\pm, u_\pm, s_\pm), \quad v_+ = v_-, s_+ = s_- \qquad \text{as} \quad x \to \pm\infty \tag{1.2}$$

[*] Received by the editors November 17, 1999; accepted for publication (in revised form) November 27, 2000; published electronically May 16, 2001.

http://www.siam.org/journals/sima/33-1/36467.html

[†] School of Political Science and Economics, Waseda University, 1-6-1 Nishi-waseda, Shinjuku, Tokyo 169-8050, Japan (kenji@mn.waseda.ac.jp). This author's work was supported in part by Waseda University grant for Special Project 98A-504 and Grant-in-Aid for Scientific Research (C)(2) 10640216 of the Ministry of Education, Science, Sports, and Culture.

[‡] Department of Mathematics, School of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku, Tokyo 169-8555, Japan (masataka@mn.waseda.ac.jp). This author's work was supported by Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists.

can be described time-asymptotically by the solution of the following equations:

$$\begin{cases} \tilde{v}_t = -\frac{1}{\alpha} p(\tilde{v}, s_0)_{xx}, \quad (t,x) \in R_+ \times R, \\ \tilde{u} = -\frac{1}{\alpha} p(\tilde{v}, s_0)_x \end{cases} \tag{1.3}$$

or

$$\begin{cases} \tilde{v}_t - \tilde{u}_x = 0, \qquad\qquad (t,x) \in R_+ \times R, \\ p(\tilde{v}, s_0)_x = -\alpha \tilde{u}. \end{cases} \tag{1.3$'$}$$

The system (1.3′) is obtained from (1.1) by approximating the momentum equation $(1.1)_2$ (the second equation of (1.1)) with Darcy's law.

In the case of isentropic flow, namely, $s(t,x) \equiv \mathit{constant}$, Hsiao and Liu [4] have proved that the solution to the Cauchy problem (1.1) converges to that of (1.3) with a rate $t^{-\frac{1}{2}}$ in the sense of the $L^2 \cap L^\infty$-norm. More precisely, for any smooth function $m_0(x)$ with compact support satisfying

$$\int_R m_0(x)dx = 1, \tag{1.4}$$

we put

$$\begin{cases} \hat{v} \equiv -\dfrac{u_+ - u_-}{\alpha} m_0(x) e^{-\alpha t}, \\ \hat{u} \equiv e^{-\alpha t} \left[u_- + (u_+ - u_-) \displaystyle\int_{-\infty}^x m_0(y)dy \right] \end{cases} \tag{1.5}$$

and uniquely determine $(\tilde{v}, \tilde{u})(t,x)$ by

$$\int_R \{v_0(x) - \tilde{v}(0,x)\}dx = \frac{u_+ - u_-}{-\alpha}. \tag{1.6}$$

Then it holds that

$$\|(v - \tilde{v} - \hat{v}, u - \tilde{u} - \hat{u})(t, \cdot)\|_{L^2 \cap L^\infty} = O(t^{-\frac{1}{2}}). \tag{1.7}$$

Moreover, the first author has obtained sharper rates than that [9]. Precisely, if we put $(v - \tilde{v} - \hat{v}, u - \tilde{u} - \hat{u}) = (V_x, z)$ due to (1.7), it holds that

$$\|(V_x, z)(t, \cdot)\|_{L^2(R)} = O(t^{-\frac{1}{2}}, t^{-1}) \tag{1.8}$$

and

$$\|(V_x, z)(t, \cdot)\|_{L^\infty(R)} = O(t^{-\frac{3}{4}}, t^{-\frac{5}{4}}), \tag{1.9}$$

which are based on the L^2-energy estimates for the reformulated problem

$$\begin{cases} V_{tt} + \{p_v(\tilde{v})V_x\}_x + \alpha V_t = \dfrac{1}{\alpha} p(\tilde{v})_{xt} - \{p(V_x + \tilde{v} + \hat{v}) - p(\tilde{v}) - p_v(\tilde{v})V_x\}_x, \\ V(0,x) = \displaystyle\int_{-\infty}^x (v - \tilde{v} - \hat{v})(0,y)dy, \ V_t(0,x) = (u - \tilde{u} - \hat{u})(0,x). \end{cases} \tag{1.10}$$

Moreover, the fact that V_{tt} decays quickly suggests that V has parabolic structure as $t \to \infty$. In fact, it is also shown that, if $v_+ = v_-$ and $(V, z)(0, x) \in L^1(R) \times L^1(R)$, then

$$\|(V_x, z)(t, \cdot)\|_{L^\infty(R)} = O(t^{-1}, t^{-\frac{3}{2}}). \tag{1.11}$$

In the nonisentropic case, the asymptotic stability in the case of $v_+ = v_-$ and $s_+ = s_-$ has been obtained in [2, 3]. Our purpose in this paper is to obtain its convergence rate, especially its second order term of asymptotic, which is on the same line as in [9, 10, 11]. See also Gallay and Raugel [1].

In the case of $v_+ \neq v_-$ and $s_+ = s_-$, Hsiao and Luo [5] have obtained the stability theorem. Furthermore, in Marcati and Pan [8], the stability results with convergence rates have been obtained in the following cases: (1) $v_+ = v_-$ and $s_+ = s_-$; (2) $p(v_-, s_-) = p(v_+, s_+)$. Hence the case $s_+ \neq s_-$ has been partly solved. We note that in these papers all data are so small that the solutions are smooth. Since large data generally yield the singularity after a finite time, we need to consider the weak solution to treat large data. See also Hsiao and Luo [6] and the references therein.

Throughout this paper we denote several generic constants by c or C. By $H^m(R)$ denote the usual Sobolev space with its norm

$$\|f\|_m := \sum_{k=0}^{m} \|\partial_x^k f\|, \quad \|\cdot\| = \|\cdot\|_0 = \|\cdot\|_{L^2(R)}.$$

2. Preliminaries and theorems. For simplicity, we restrict our case to $v_+ = v_- := \underline{v}$, $u_+ = u_- := 0$, $s_+ = s_- := \underline{s}$, so that $(\hat{v}, \hat{u})(t, x) \equiv (0, 0)$. The constant α is normalized to 1 without loss of generality.

First, let us consider the problem (1.3) in order to reformulate our problem (1.1), (1.2). Since $\tilde{u}$ is defined by $(1.3)_2$, we investigate the Cauchy problem of $\tilde{v}$ to the parabolic equation

$$\begin{cases} \tilde{v}_t + p(\tilde{v}, s_0)_{xx} = 0, & (t, x) \in R_+ \times R, \\ \tilde{v}(0, x) = \tilde{v}_0(x) \to \underline{v} & (x \to \pm\infty). \end{cases} \tag{2.1}$$

Equation $(2.1)_1$ has a stationary solution $\bar{v}(x)$ defined by

$$p(\bar{v}(x), s_0(x)) = p(\underline{v}, \underline{s}). \tag{2.2}$$

For a typical form of $p(v, s)$ in gas dynamics, $\bar{v}$ is given by

$$\bar{v}(x) = e^{\frac{1}{\gamma}(s_0(x) - \underline{s})}\underline{v}. \tag{2.3}$$

Our first proposition is on the property of $\tilde{v}$, which is necessary to investigate the behavior of solutions to (1.1), (1.2).

Proposition 2.1 (asymptotic property of $\tilde{v}$). *Suppose that $p(v, s)$ is a smooth function with $p > 0, p_v < 0$ for $v > 0$. If $(\tilde{v}_0 - \bar{v}, s_0 - \underline{s}) \in H^6(R) \times H^6(R)$ is sufficiently small, then there exists a unique global solution $\tilde{v}(t, x)$ to* (2.1)*, which satisfies that*

$$\partial_t^i(\tilde{v} - \bar{v}) \in C([0, \infty); H^{6-2i}(R)), \partial_t^j q_x \in C([0, \infty); H^{5-2j}(R)),$$

$$q_{tttx} \in L^2(0, \infty; L^2(R)),$$

where $q = p(\bar{v}, s) - p(\tilde{v}, s)$ *and* $i = 0, 1, 2, 3,\ j = 0, 1, 2,$ *and that*

$$\begin{aligned}&\sum_{k=0}^{3}(1+t)^{2k}\|\partial_t^k(\tilde{v}-\bar{v})(t)\|^2 + \sum_{k=0}^{2}(1+t)^{2k+1}\|\partial_t^k q_x(t)\|^2 \\ &+ \int_0^t \left(\sum_{k=1}^{3}(1+\tau)^{2k-1}\|\partial_t^k(\tilde{v}-\bar{v})(\tau)\|^2 + \sum_{k=0}^{3}(1+\tau)^{2k}\|\partial_t^k q_x(\tau)\|^2\right) d\tau \\ &\le C\|(\tilde{v}_0(\cdot)-\bar{v}(\cdot), s_0(\cdot)-\underline{s})\|_6^2.\end{aligned} \tag{2.4}$$

Moreover, if $(\tilde{v}_0 - \bar{v}, s_0 - \underline{s}) \in L^1(R) \times L^1(R)$ *is assumed, then* $\tilde{v}$ *satisfies*

$$\begin{cases} \|\tilde{v}(t,\cdot) - (\bar{v}(\cdot) + \bar{\theta}_0(t,\cdot))\|_{L^\infty} \le C(1+t)^{-1}, \\ \|(\tilde{v} - \bar{\theta}_0)_t(t,\cdot)\|_{L^\infty} \le C(1+t)^{-2}, \\ \|(\tilde{v} - \bar{\theta}_0)_{tt}(t,\cdot)\|_{L^\infty} \le C(1+t)^{-\frac{11}{4}}, \\ \|\tilde{v}(t,\cdot) - (\bar{v}(\cdot) + \bar{\theta}_0(t,\cdot))\|_{L^2} \le C(1+t)^{-\frac{3}{4}}, \\ \|(\tilde{v} - \bar{\theta}_0)_t(t,\cdot)\|_{L^2} \le C(1+t)^{-\frac{7}{4}}, \\ \|(\tilde{v} - \bar{\theta}_0)_{tt}(t,\cdot)\|_{L^2} \le C(1+t)^{-\frac{5}{2}}, \end{cases} \tag{2.5}$$

where $\bar{\theta}_0$ *is given by the explicit formula*

$$\bar{\theta}_0(t,x) = \frac{-p_v(\underline{v},\underline{s})}{-p_v(\bar{v},s)} \int_R G(t,x-y)\{\tilde{v}_0(y) - \bar{v}(y)\}dy \tag{2.6}$$

by using the Green function of $v_t + p_v(\underline{v},\underline{s})v_{xx} = 0$.

Remark 1. In Proposition 2.1, we assume that the initial disturbance is in $H^6(R)$ in order to obtain the decay estimates for $\tilde{v}_{ttt},\ q_{tttx}$, etc., which will be used in the proofs of Theorems 1 and 2 below. The function $\bar{\theta}_0(t,x)$ satisfies

$$(\bar{\theta}_0)_t = \frac{\underline{a}}{a(x)}(a(x)\bar{\theta}_0)_{xx}, \tag{2.7}$$

where $a(x) = -p_v(\bar{v}(x), s_0(x))$ and $\underline{a} = -p_v(\underline{v},\underline{s})$. Since $\|(a(\cdot)\bar{\theta}_0, a(\cdot)\bar{\theta}_{0t})(t)\|_{L^\infty} = O(t^{-\frac{1}{2}}, t^{-\frac{3}{2}})$, etc., we can say from (2.5) that $\bar{v}(x) + \bar{\theta}_0(t,x)$ is an asymptotic profile of $\tilde{v}$ as $t \to \infty$. It seems to be curious, because $\bar{\theta}_0$ satisfies (2.7) instead of $(\bar{\theta}_0)_t = (a(x)\bar{\theta}_0)_{xx}$, the linearized equation of (2.1) around $\tilde{v}$. However, we have adopted $\bar{\theta}_0$ in (2.6) which has an explicit formula.

We also obtain the asymptotic property of $\tilde{u}$.

PROPOSITION 2.2. *The function* $\tilde{u}$ *defined by* $(1.3')_2$ *for* $\tilde{v}$ *obtained in Proposition* 2.1 *satisfies*

$$\begin{cases} \|(\tilde{u} - \bar{q}_{0x})(t,\cdot)\|_{L^\infty} \le C(1+t)^{-\frac{3}{2}}\log(2+t), \\ \|(\tilde{u} - \bar{q}_{0x})(t,\cdot)\|_{L^2} \le C(1+t)^{-\frac{5}{4}}, \end{cases} \tag{2.8}$$

where $\bar{q}_0$ *is given by*

$$\bar{q}_0(t,x) = -p_v(\underline{v},\underline{s}) \int_R G(t,x-y)\{\tilde{v}_0(y) - \bar{v}(y)\}dy. \tag{2.9}$$

The proofs of Propositions 2.1 and 2.2 will be given in section 3.

We now turn to the original problem (1.1), (1.2) with $v_\pm = \underline{v}$ and $u_\pm = 0$. If $\tilde{v}_0(x) = \underline{v} + \frac{\delta_0}{\sqrt{4\pi}} \exp(-\frac{(x-x_0)^2}{4})$, $\delta_0 = \int_R (v_0(y) - \underline{v})dy$, then $\int_R (v_0 - \tilde{v}_0)(x)dx = 0$. Hence $\int_R (v - \tilde{v})(t,x)dx \equiv 0$ follows from $(1.1)_1$ and $(1.3')_1$. Thus, putting

$$(V,z)(t,x) = \left(\int_{-\infty}^x (v-\tilde{v})(t,y)dy, (u-\tilde{u})(t,x) \right), \tag{2.10}$$

we have the reformulated problem

$$(RP) \qquad \begin{cases} V_t - z = 0, \\ z_t + \{p(V_x + \tilde{v}, s) - p(\tilde{v}, s)\}_x + z = p(\tilde{v}, s)_{xt}, \\ (V,z)(0,x) = (V_0, z_0)(x) \\ \qquad\qquad := \left(\int_{-\infty}^x \{v_0(y) - \tilde{v}(0,y)\}dy, u_0(x) - \tilde{u}(0,x) \right) \\ \qquad\qquad \to 0 \qquad \text{as} \qquad x \to \pm\infty, \end{cases}$$

where $s(t,x) \equiv s_0(x) =: s(x)$.

Hsiao and Serre [2, 3] have shown that $(\tilde{v}, \tilde{u})(t,x) \to (\bar{v}(x), 0)$ as $t \to \infty$ and that $(V_x, z) = (v - \tilde{v}, u - \tilde{u})(t,x) \to 0$ as $t \to \infty$ under suitable smallness conditions. Namely, the solution (v,u) to (1.1), (1.2) tends to $(\bar{v}(x), 0)$ as t tends to infinity. In this paper, we obtain those convergence rates by applying not only the L^2-energy method but also the Green function of parabolic equation.

Using the property of $\tilde{v}$ in Proposition 2.1, we obtain the following first main theorem based on the L^2-energy method.

Theorem 1. *In addition to the assumptions in Proposition* 2.1*, suppose that* $v_0 - \tilde{v}_0 \in L^1(R)$. *If* $(V_0, z_0) \in H^3(R) \times H^2(R)$ *is sufficiently small, then there exists a unique global solution* $(V,z)(t,x)$ *to* (RP)*, which satisfies*

$$\begin{aligned} &\sum_{k=0}^{1} (1+t)^k \|\partial_x^k V(t)\|^2 + \sum_{k=2}^{3} (1+t)^k \|\partial_x^{k-1} P(t)\|^2 + \sum_{k=0}^{1} (1+t)^{k+2} \|\partial_x^k z(t)\|^2 \\ &+ (1+t)^3 \|z_{xx}(t)\|^2 + (1+t)^4 \|P_{xt}(t)\|^2 + \sum_{k=0}^{1} (1+t)^{2k+3} \|\partial_x^k z_t(t)\|^2 + (1+t)^5 \|z_{tt}(t)\|^2 \\ &+ \int_0^t \left\{ \|V_x(\tau)\|^2 + (1+\tau)\|P_x(\tau)\|^2 + (1+\tau)^2 \|P_{xx}(\tau)\|^2 + (1+\tau)^3 \|P_{xt}(\tau)\|^2 \right\} d\tau \\ &+ \int_0^t \left\{ \sum_{k=0}^{2} (1+\tau)^{k+1} \|\partial_x^k z(\tau)\|^2 + \sum_{k=0}^{1} (1+\tau)^{k+3} \|\partial_t^k z_t(\tau)\|^2 + (1+\tau)^5 \|z_{tt}(\tau)\|^2 \right\} d\tau \\ &\le C \left(\|V_0\|_3^2 + \|z_0\|_2^2 + \|(\tilde{v}_0(\cdot) - \bar{v}(\cdot), s_0(\cdot) - \underline{s})\|_6^2 \right). \end{aligned} \tag{2.11}$$

Here $P = P(V_x) = P(V_x; \tilde{v}, s)(t,x) := p(V_x(t,x) + \tilde{v}(t,x), s(x)) - p(\tilde{v}(t,x), s(x))$.

Proof. The proof is given by the same method as in [9]. □

Remark 2. The decay estimate of (V_x, z), corresponding with (1.9), is derived as follows. By the Sobolev inequality and $C^{-1}|V_x| \le |P(V_x)| \le C|V_x|$,

$$\begin{cases} \|V_x(t,\cdot)\|_{L^\infty} \le C\|P(t;V_x,\tilde{v},s)\| \\ \qquad\qquad\qquad \le C\|P(t;V_x,\tilde{v},s)\|^{\frac{1}{2}}\|P(t;V_x,\tilde{v},s)_x\|^{\frac{1}{2}} \\ \qquad\qquad\qquad \le C\|V_x(t)\|^{\frac{1}{2}}\|P(t;V_x,\tilde{v},s)_x\|^{\frac{1}{2}} \\ \qquad\qquad\qquad \le C(1+t)^{-\frac{1}{4}}(1+t)^{-\frac{1}{2}} = C(1+t)^{-\frac{3}{4}} \end{cases}$$

and

$$\begin{aligned} \|z(t,\cdot)\|_{L^\infty} &\le C\|z(t)\|^{\frac{1}{2}}\|z_x(t)\|^{\frac{1}{2}} \\ &\le C(1+t)^{-\frac{1}{2}}(1+t)^{-\frac{3}{4}} = C(1+t)^{-\frac{5}{4}}. \qquad \square \end{aligned}$$

Next, we obtain the optimal convergence rate, corresponding with (1.11), assuming that $(V_0, z_0) \in L^1(R) \times L^1(R)$. The linearized problem of (RP) is

$$\begin{cases} V_{tt} + \{p_v(\tilde{v},s)V_x\}_x + V_t = p(\tilde{v},s)_{xt} - F_x, \\ V(0,x) = V_0(x), V_t(0,x) = z_0(x), \end{cases} \tag{2.12}$$

where $F = p(V_x + \tilde{v}, s) - p(\tilde{v}, s) - p_v(\tilde{v}, s)V_x$. Regarding (2.12) as the parabolic equation of V with "forcing terms," we have the expression

$$\begin{aligned} V(t,x) = &\int_R G(t,x-y)V_0(y)dy - \int_0^t\int_R G(t-\tau,x-y)(V_{tt}+F_x)dyd\tau \\ &+ \int_0^t\int_R G(t-\tau,x-y)p(\tilde{v},s)_{xt}dyd\tau \\ &- \int_0^t\int_R G(t-\tau,x-y)\{(p_v(\tilde{v},s)-p_v(\underline{v},\underline{s}))V_x\}_x dyd\tau, \end{aligned} \tag{2.13}$$

which are estimated by using the result in Theorem 1. Thus we have the following theorem.

THEOREM 2. *In addition to the assumptions in Theorem* 1, *suppose that* $(V_0, z_0) \in L^1(R) \times L^1(R)$. *Then the global solution* (V, z) *to* (RP) *satisfies*

$$\begin{cases} \|(V_x, z)(t)\|_{L^\infty} = O(t^{-1}, t^{-\frac{3}{2}}), \\ \|(V_x, z)(t)\|_{L^2} = O(t^{-\frac{3}{4}}, t^{-\frac{5}{4}}). \end{cases} \tag{2.14}$$

Combining Propositions 2.1 and 2.2 and Theorem 2, we have the last theorem.

THEOREM 3. *Suppose the same assumptions as those in Theorem* 2. *Then the global solution to* (1.1), (1.2) *satisfies*

$$\begin{cases} \|v(t,\cdot) - (\bar{v}(\cdot) + \bar{\theta}_0(t,\cdot))\|_{L^\infty} \le C(1+t)^{-1}, \\ \|u(t,\cdot) - \bar{q}_{0x}(t,\cdot)\|_{L^\infty} \le C(1+t)^{-\frac{3}{2}}\log(2+t), \\ \|v(t,\cdot) - (\bar{v}(\cdot) + \bar{\theta}_0(t,\cdot))\|_{L^2} \le C(1+t)^{-\frac{3}{4}}, \\ \|u(t,\cdot) - \bar{q}_{0x}(t,\cdot)\|_{L^2} \le C(1+t)^{-\frac{5}{4}}. \end{cases} \tag{2.15}$$

Remark 3. If, eventually, $\int_R (v_0 - \bar{v})(x)dx = 0$, then we can put $(V, z)(t, x) = (\int_{-\infty}^{x}(v(t, y) - \bar{v}(y))dy, (u - \bar{u})(t, x))$, which yields the simpler problem

$$\begin{cases} V_t - z = 0, \\ z_t + \{p(V_x + \bar{v}, s) - p(\bar{v}, s)\}_x + z = 0. \end{cases}$$

In the result, we have $\|v(t, \cdot) - \bar{v}(\cdot)\|_{L^\infty} = O(t^{-1})$ instead of (2.15). Hence, the estimate (2.15) almost implies that the diffusion wave $\bar{\theta}_0(t, x)$ carries on the amount $\int_R (v_0(x) - \bar{v}(x))dx$. In (2.15), we could not remove $\log(2 + t)$.

The proofs of Theorems 1–3 will be done in sections 4–5.

3. Asymptotic behavior of the parabolic equation. In this section, we prove Propositions 2.1 and 2.2. By setting θ as

$$\theta := \tilde{v} - \bar{v}, \tag{3.1}$$

the Cauchy problem (2.1) is rewritten as

$$\begin{cases} \theta_t = q(\theta, x)_{xx}, \\ \theta|_{t=0} = \theta_0(x) \equiv \tilde{v}_0(x) - \bar{v}(x), \end{cases} \tag{3.2}$$

where

$$q(\theta, x) \equiv p(\bar{v}(x), s(x)) - p(\theta + \bar{v}(x), s(x)). \tag{3.3}$$

Applying the L^2-energy method, we first prove the following proposition.

PROPOSITION 3.1. *Suppose that $p(v, s)$ is a smooth function with $p > 0, p_v < 0$ for $v > 0$. If $(\theta_0, s_0 - \underline{s}) \in H^6(R) \times H^6(R)$ is sufficiently small, then there exists a unique global solution $\theta(t, x)$ to* (3.2)*, which satisfies that*

$$\begin{aligned} &\partial_t^i \theta \in C([0, \infty); H^{6-2i}(R)), \partial_t^j q_x \in C([0, \infty); H^{5-2j}(R)), \\ &\quad q_{tttx} \in L^2(0, \infty; L^2(R)), \end{aligned}$$

where $i = 0, 1, 2, 3$ and $j = 0, 1, 2$, and that

$$\begin{aligned} &\sum_{k=0}^{3}(1 + t)^{2k}\|\partial_t^k \theta(t)\|^2 + \sum_{k=0}^{2}(1 + t)^{2k+1}\|\partial_t^k q_x(t)\|^2 \\ &+ \int_0^t \left(\sum_{k=1}^{3}(1 + \tau)^{2k-1}\|\partial_t^k \theta(\tau)\|^2 + \sum_{k=0}^{3}(1 + \tau)^{2k}\|\partial_t^k q_x(\tau)\|^2\right) d\tau \\ &\leq C\|(\theta_0(\cdot), s_0(\cdot) - \underline{s})\|_6^2. \end{aligned} \tag{3.4}$$

Next, using the Green function, we obtain an asymptotic profile under the assumption of $\theta_0 \in L^1(R)$, which gives the optimal decay rates of θ.

PROPOSITION 3.2. *In addition to the assumptions in Proposition* 3.1*, suppose* *that* $(\theta_0, s_0 - \underline{s}) \in L^1(R) \times L^1(R)$. *Then the global solution* $\theta(t,x)$ *to* (3.2) *satisfies*

$$\begin{cases} \|(\theta - \bar{\theta}_0)(t,\cdot)\|_{L^\infty} \leq C(1+t)^{-1}, \\ \|(\theta - \bar{\theta}_0)_t(t,\cdot)\|_{L^\infty} \leq C(1+t)^{-2}, \\ \|(\theta - \bar{\theta}_0)_{tt}(t,\cdot)\|_{L^\infty} \leq C(1+t)^{-\frac{11}{4}}, \\ \|(\theta - \bar{\theta}_0)(t,\cdot)\|_{L^2} \leq C(1+t)^{-\frac{3}{4}}, \\ \|(\theta - \bar{\theta}_0)_t(t,\cdot)\|_{L^2} \leq C(1+t)^{-\frac{7}{4}}, \\ \|(\theta - \bar{\theta}_0)_{tt}(t,\cdot)\|_{L^2} \leq C(1+t)^{-\frac{5}{2}}. \end{cases} \tag{3.5}$$

Since

$$\begin{aligned} \tilde{u} &= -p(\theta + \bar{v}, s_0)_x = -\{p(\theta + \bar{v}, s_0) - p(\underline{v}, \underline{s})\}_x \\ &= -\{p(\theta + \bar{v}, s_0) - p(\bar{v}, s_0)\}_x = q_x, \end{aligned}$$

we also estimate $\tilde{u}$ by using the Green function. Differentiating q in t, we obtain $q_t = -p_v(\theta + \bar{v}, s_0)\theta_t$. Substituting this into (3.2), we have

$$q_t = -p_v(\theta + \bar{v}, s_0) q_{xx}. \tag{3.6}$$

Then the following proposition holds.

PROPOSITION 3.3. *Suppose the same assumptions as those in Proposition* 2.2. *Then the global solution* $q(t,x)$ *of* (3.6) *with* $\tilde{u} = q_x$ *satisfies*

$$\begin{cases} \|(q - \bar{q}_0)(t,\cdot)\|_{L^\infty} \leq C(1+t)^{-1}, \\ \|(q - \tilde{q}_0)_x(t,\cdot)\|_{L^\infty} \leq C(1+t)^{-\frac{3}{2}} \log(2+t), \\ \|(q - \bar{q}_0)(t,\cdot)\|_{L^2} \leq C(1+t)^{-\frac{3}{4}}, \\ \|(q - \tilde{q}_0)_x(t,\cdot)\|_{L^2} \leq C(1+t)^{-\frac{5}{4}}. \end{cases} \tag{3.7}$$

The assertions of Proposition 2.1 follow from Proposition 3.1 and 3.2, and those of Proposition 2.2 follow from Proposition 3.3. The proofs of Propositions 3.1 and 3.2 will be divided into several steps. Proposition 3.3 will be proved at the end of this section.

Proof of Proposition 3.1. The proof is given by the combination of the local existence with a priori estimates. Since the local existence theorem is obtained in a standard way [7], we devote ourselves to the estimates under the a priori assumption

$$N_1(T) := \sup_{0 \leq t \leq T} \left\{ \sum_{k=0}^{3} (1+t)^k \|\partial_t^k \theta(t)\| + \sum_{k=0}^{2} (1+t)^{k+\frac{1}{2}} \|\partial_t^k q_x(t)\| \right\} \leq \varepsilon. \tag{3.8}$$

Estimate 1. Multiplying $(3.2)_1$ by $q(\theta, x)$ and integrating it over $[0,t] \times R$, we get

$$\begin{aligned} &\int_R Q(\theta, x) dx + \int_0^t \int_R \{q(\theta, x)_x\}^2 dx d\tau \\ &\qquad = \int_R Q(\theta, x) dx \bigg|_{t=0} \leq C \|(\theta_0, s_0 - \underline{s})\|^2, \end{aligned} \tag{3.9}$$

where $Q(\theta, x) = \int_0^\theta q(\eta, x)d\eta$, which is equivalent to θ^2.

Estimate 2. Define $\bar{V} = q(\theta, x)_x$. Then

$$\bar{V}_t = q_{xt}. \tag{3.10}$$

Mutiplying (3.10) by $(1+t)\bar{V}$ and integrating it over R, we have

$$\frac{1}{2}\frac{d}{dt}\left\{(1+t)\int_R q_x^2 dx\right\} + (1+t)\int_R (-p_v)\theta_t^2 dx = \frac{1}{2}\int_R q_x^2 dx. \tag{3.11}$$

Integrating (3.11) and applying (3.9), we get

$$\begin{aligned} &(1+t)\int_R q_x^2 dx + \int_0^t (1+\tau)\int_R (-p_v)\theta_t^2 dx d\tau \\ &\le C\|(\theta_0, s_0 - \underline{s})\|_1^2. \end{aligned} \tag{3.12}$$

Estimate 3. Differentiate $(3.2)_1$ in t:

$$(\theta_t)_t = \{q(\theta, x)_t\}_{xx}. \tag{3.13}$$

Multiplying (3.13) by $(1+t)^2 q_t$ and integrating it over R, we have

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\left\{(1+t)^2\int_R (-p_v)\theta_t^2 dx\right\} + (1+t)^2\int_R q_{tx}^2 dx \\ &= (1+t)\int_R (-p_v)\theta_t^2 dx - (1+t)^2\int_R (-p_v)_t\theta_t^2 dx \\ &\le (1+t)\int_R (-p_v)\theta_t^2 dx + C(1+t)^2 \sup|\theta_t| \int_R \theta_t^2 dx, \end{aligned} \tag{3.14}$$

and hence, by (3.12) and $N_1(T) \le \varepsilon$,

$$\begin{aligned} &(1+t)^2\int_R (-p_v)\theta_t^2 dx + \int_0^t (1+\tau)^2\int_R q_{tx}^2 dx d\tau \\ &\le C\|(\theta_0, s_0 - \underline{s})\|_2^2. \end{aligned} \tag{3.15}$$

Estimate 4. Multiply (3.13) by $-p_v(\theta + \bar{v}, s)$ to obtain

$$q_{tt} = (-p_v)q_{txx} + (-p_v)_t\theta_t. \tag{3.16}$$

Multiplying (3.16) by $(1+t)^3(-q_{txx})$ yields

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\left\{(1+t)^3\int_R q_{tx}^2 dx\right\} + (1+t)^3\int_R (-p_v)q_{txx}^2 dx \\ &= \frac{3}{2}(1+t)^2\int_R q_{tx}^2 dx + (1+t)^3\int_R q_{txx}(p_v)_t\theta_t dx \end{aligned} \tag{3.17}$$

and

$$\begin{aligned} &(1+t)^3\int_R q_{tx}^2 dx + \int_0^t (1+\tau)^3\int_R (-p_v)q_{txx}^2 dx d\tau \\ &\le C\|(\theta_0, s_0 - \underline{s})\|_3^2 \end{aligned} \tag{3.18}$$

in a similar fashion to Estimate 3.

Estimate 5. Similar methods to Estimates 3 and 4 give the estimates

$$\begin{aligned}&(1+t)^4\int_R(-p_v)\theta_{tt}^2dx+\int_0^t(1+\tau)^4\int_R q_{ttx}^2dxd\tau\\&\quad\le C\|(\theta_0,s_0-\underline{s})\|_4^2,\end{aligned}\tag{3.19}$$

$$\begin{aligned}&(1+t)^5\int_R q_{ttx}^2dx+\int_0^t(1+\tau)^5\int_R(-p_v)q_{ttxx}^2dxd\tau\\&\quad\le C\|(\theta_0,s_0-\underline{s})\|_5^2,\end{aligned}\tag{3.20}$$

and

$$\begin{aligned}&(1+t)^6\int_R(-p_v)\theta_{ttt}^2dx+\int_0^t(1+\tau)^6\int_R q_{tttx}^2dxd\tau\\&\quad\le C\|(\theta_0,s_0-\underline{s})\|_6^2.\end{aligned}\tag{3.21}$$

Combining Estimates 1–5 completes the proof of Proposition 3.1. □

Remark 4. Since $q(\theta,x)=p(\bar{v},s)-p(\theta+\bar{v},s)=-p_v(\cdot,s)\theta$, the Sobolev inequality and (3.4) yield

$$\begin{aligned}\sup_R|\theta|&\le C\sup_R|q|\\&\le C\|q(t)\|^{\frac12}\|q_x(t)\|^{\frac12}\\&\le C\|\theta(t)\|^{\frac12}\|q_x(t)\|^{\frac12}\le C(1+t)^{-\frac14}.\end{aligned}\tag{3.22}$$

Due to $q_t=-p_v(\theta+\bar{v},s)\theta_t$, we have

$$\begin{aligned}\sup_R|\theta_t|&\le C\sup_R|q_t|\\&\le C\|q_t(t)\|^{\frac12}\|q_{tx}(t)\|^{\frac12}\\&\le C\|\theta_t(t)\|^{\frac12}\|q_{tx}(t)\|^{\frac12}\le C(1+t)^{-\frac54}.\end{aligned}\tag{3.23}$$

Since $q_{tt}=-p_{vv}\theta_t^2-p_v\theta_{tt}$,

$$\begin{aligned}\sup\nolimits_R|q_{tt}|&\le C\|q_{tt}(t)\|^{\frac12}\|q_{ttx}(t)\|^{\frac12}\\&\le C\left(\|\theta_t^2(t)\|+\|\theta_{tt}(t)\|\right)^{\frac12}\|q_{ttx}(t)\|^{\frac12}\\&\le C\left(\|\theta_t(t)\|_{L^\infty}\|\theta_t(t)\|+\|\theta_{tt}(t)\|\right)^{\frac12}\|q_{ttx}(t)\|^{\frac12}\\&\le C(1+t)^{-\frac94}\end{aligned}\tag{3.24}$$

and

$$\sup_R|\theta_{tt}|\le C\left(\sup_R|q_{tt}|+\sup_R|\theta_t(t)|^2\right).\tag{3.25}$$

Therefore,

$$\sup_R |\theta_{tt}(t)| \le C(1+t)^{-\frac{9}{4}}. \tag{3.26}$$

Proof of Proposition 3.2.

First step. We first investigate the Cauchy problem for the homogeneous linearized equation to $(3.2)_1$:

$$\begin{cases} \bar{\theta}_t = (a(x)\bar{\theta})_{xx}, \\ \bar{\theta}|_{t=0} = \theta_0(x), \end{cases} \tag{3.27}$$

where $a(x) \equiv -p_v(\bar{v}(x), s(x)) \to -p_v(\underline{v}, \underline{s}) \equiv \underline{a}$ as $x \to \pm\infty$. To obtain the precise decay estimates of $\bar{\theta}$, we here again combine the L^2-energy method with the explicit formula using the Green function. First, multiplying $(3.27)_1$ by $a(x)\bar{\theta}$ and integrating it over $[0,t] \times R$, we have

$$\frac{1}{2}\int_R a(x)\bar{\theta}^2 dx + \int_0^t \int_R \{(a(x)\bar{\theta})_x\}^2 dx d\tau = \frac{1}{2}\int_R a(x)\bar{\theta}_0^2 dx. \tag{3.28}$$

Next, multiplying $(3.27)_1$ by $a(x)$ and differentiating it in x, we obtain

$$(a(x)\bar{\theta})_{xt} = \{a(x)(a(x)\bar{\theta})_{xx}\}_x. \tag{3.29}$$

Multiplying (3.29) by $(1+t)(a(x)\bar{\theta})_x$ and integrating it over $[0,t] \times R$, we have

$$\begin{aligned} &\frac{1}{2}(1+t)\int_R \{(a(x)\bar{\theta})_x\}^2 dx + \int_0^t (1+\tau)\int_R a(x)\{(a(x)\bar{\theta})_{xx}\}^2 dx d\tau \\ &\quad = \frac{1}{2}\int_R \{(a(x)\bar{\theta}_0)_x\}^2 dx + \int_0^t \int_R \{(a(x)\bar{\theta})_x\}^2 dx d\tau. \end{aligned} \tag{3.30}$$

By virtue of (3.28) and $(a(x)\bar{\theta})_{xx} = \bar{\theta}_t$, we obtain

$$\begin{aligned} &(1+t)\int_R \{(a(x)\bar{\theta})_x\}^2 dx + \int_0^t (1+\tau)\int_R a(x)\left[\{(a(x)\bar{\theta})_{xx}\}^2 + \bar{\theta}_t^2\right] dx d\tau \\ &\quad \le C\|(\theta_0, s_0 - \underline{s})\|_1^2. \end{aligned} \tag{3.31}$$

Since $(\partial_t^k \bar{\theta})_t = (a(x)\partial_t^k \bar{\theta})_{xx}, k = 1, 2, 3,$ similar estimates to those above give the following.

Lemma 3.1. *If $(\theta_0, s - \underline{s}) \in H^6(R) \times H^6(R)$, then it holds that*

$$\begin{aligned} &\sum_{k=0}^{3}\left\{(1+t)^{2k}\int_R a(x)(\partial_t^k \bar{\theta})^2 dx + (1+t)^{2k+1}\int_R \{(a(x)\partial_t^k \bar{\theta})_x\}^2 dx\right\} \\ &\quad + \sum_{k=0}^{2}\int_0^t (1+\tau)^{2k}\int_R \{a(x)(\partial_t^k \bar{\theta})_x\}^2 dx d\tau \\ &\quad + \sum_{k=0}^{2}\int_0^t (1+\tau)^{2k+1}\int_R a(x)[\{(a(x)\partial_t^k \bar{\theta})_{xx}\}^2 + (\partial_t^k \bar{\theta}_t)^2] dx d\tau \\ &\quad \le C\|(\theta_0, s_0 - \underline{s})\|_6^2. \end{aligned} \tag{3.32}$$

Remark 5. By the Sobolev inequality and (3.32), we obtain

$$\begin{aligned}\sup_R|\bar{\theta}| &= \sup_R |a(x)^{-1}| \sup_R |a(x)\bar{\theta}| \\ &\le C\|\sqrt{a(x)}\bar{\theta}(t)\|^{\frac{1}{2}}\|(a(x)\bar{\theta})_x(t)\|^{\frac{1}{2}} \\ &\le C(1+t)^{-\frac{1}{4}},\end{aligned} \tag{3.33}$$

$$\begin{aligned}\sup_R|\bar{\theta}_t| &\le C\|\sqrt{a(x)}\bar{\theta}_t(t)\|^{\frac{1}{2}}\|(a(x)\bar{\theta}_t)_x(t)\|^{\frac{1}{2}} \\ &\le C(1+t)^{-\frac{5}{4}},\end{aligned} \tag{3.34}$$

and

$$\begin{aligned}\sup_R|\bar{\theta}_{tt}| &\le C\|\sqrt{a(x)}\bar{\theta}_{tt}(t)\|^{\frac{1}{2}}\|(a(x)\bar{\theta}_{tt})_x(t)\|^{\frac{1}{2}} \\ &\le C(1+t)^{-\frac{9}{4}}.\end{aligned} \tag{3.35}$$

Second step. Assuming that $\theta_0 \in L^1(R)$, we now obtain an asymptotic profile $\bar{\theta}_0$ of $\bar{\theta}$ defined in (2.6). Rewrite $(3.27)_1$ as

$$\bar{\theta}_t = \underline{a}\bar{\theta}_{xx} + \{(a(x)-\underline{a})\bar{\theta}\}_{xx}$$

to have the expression

(3.36)

$$\bar{\theta}(t,x) = \int_R G(t,x-y)\theta_0(y)dy + \int_0^t\int_R G(t-\tau,x-y)\{(a(y)-\underline{a})\bar{\theta}(\tau,y)\}_{yy}dyd\tau,$$

where

$$G(t,x) = \frac{1}{\sqrt{4\pi\underline{a}t}}\exp\left(-\frac{x^2}{4\underline{a}t}\right).$$

Integration by parts yields

$$\begin{aligned}&\int_{\frac{t}{2}}^t\int_R G\cdot\{(a(y)-\underline{a})\bar{\theta}(\tau,y)\}_{yy}dyd\tau \\ &\quad= -\frac{1}{\underline{a}}\int_{\frac{t}{2}}^t\int_R G_\tau\cdot(a(y)-\underline{a})\bar{\theta}(\tau,y)dyd\tau \\ &\quad= -\frac{1}{\underline{a}}\left[\int_R G\cdot(a(y)-\underline{a})\bar{\theta}(\tau,y)dy\right]_{\frac{t}{2}}^t + \frac{1}{\underline{a}}\int_{\frac{t}{2}}^t\int_R G\cdot(a(y)-\underline{a})\bar{\theta}_\tau(\tau,y)dyd\tau \\ &\quad= -\frac{1}{\underline{a}}(a(x)-\underline{a})\bar{\theta}(t,x) + \frac{1}{\underline{a}}\int_R G\left(\frac{t}{2},x-y\right)(a(y)-\underline{a})\bar{\theta}(\tau,y)dy \\ &\qquad+\frac{1}{\underline{a}}\int_{\frac{t}{2}}^t\int_R G\cdot(a(y)-\underline{a})\bar{\theta}_\tau(\tau,y)dyd\tau.\end{aligned}$$

Hence by (3.36)

$$
\begin{aligned}
(\bar\theta-\bar\theta_0)(t,x) &= \frac{1}{a(x)}\int_R G\left(\frac{t}{2},x-y\right)(a(y)-\underline{a})\bar\theta\left(\frac{t}{2},y\right)dy \\
&\quad+\frac{\underline{a}}{a(x)}\int_0^{\frac{t}{2}}\int_R G_{yy}(t-\tau,x-y)(a(y)-\underline{a})\bar\theta(\tau,y)dyd\tau \\
&\quad+\frac{1}{a(x)}\int_{\frac{t}{2}}^{t}\int_R G(t-\tau,x-y)(a(y)-\underline{a})\bar\theta_\tau(\tau,y)dyd\tau \\
&\equiv I_1+I_2+I_3.
\end{aligned}
\tag{3.37}
$$

Here we have used $\bar\theta_0=\frac{\underline{a}}{a(x)}\int_R G\cdot\theta_0(y)dy$.

By $\theta_0\in L^1(R)$, it is easy to see

$$
|\bar\theta_0(t,x)|\le Ct^{-\frac{1}{2}}. \tag{3.38}
$$

Since $s_0-\underline{s}\in L^1(R)$, $a(y)-\underline{a}\in L^1(R)$. Hence

$$
\begin{aligned}
|I_1| &\le C\sup_R\left|G\left(\frac{t}{2},x-y\right)\right|\sup_R\left|\bar\theta\left(\frac{t}{2},y\right)\right|\|a(\cdot)-\underline{a}\|_{L^1} \\
&\le Ct^{-\frac{3}{4}},
\end{aligned}
\tag{3.39}
$$

$$
\begin{aligned}
|I_2| &\le C\int_0^{\frac{t}{2}}\sup_R|G_{yy}|\sup_R|\bar\theta(\tau,y)|\|a(\cdot)-\underline{a}\|_{L^1}d\tau \\
&\le C\int_0^{\frac{t}{2}}(t-\tau)^{-\frac{3}{2}}(1+\tau)^{-\frac{1}{4}}d\tau, \\
&\le Ct^{-\frac{3}{4}},
\end{aligned}
\tag{3.40}
$$

and

$$
\begin{aligned}
|I_3| &\le C\int_{\frac{t}{2}}^{t}\sup_R|G|\sup_R|\bar\theta_\tau(\tau,y)|\|a(\cdot)-\underline{a}\|_{L^1}d\tau \\
&\le C\int_0^{\frac{t}{2}}(t-\tau)^{-\frac{1}{2}}(1+\tau)^{-\frac{5}{4}}d\tau \\
&\le Ct^{-\frac{3}{4}}.
\end{aligned}
\tag{3.41}
$$

Here we have used (3.33) and (3.34). Combining (3.37)–(3.41), we obtain

$$
\sup_R|(\bar\theta-\bar\theta_0)(t,x)|\le Ct^{-\frac{3}{4}}. \tag{3.42}
$$

Next, we estimate $\sup_R |(\bar\theta - \bar\theta_0)_t(t,x)|$. Differentiate (3.37) in t to have

$$\begin{aligned}(\bar\theta-\bar\theta_0)_t(t,x) &= \frac{1}{a(x)}\int_R G_t\left(\frac{t}{2},x-y\right)(a(y)-\underline{a})\bar\theta\left(\frac{t}{2},y\right)dy\\ &+\frac{1}{a(x)}\int_R G\left(\frac{t}{2},x-y\right)(a(y)-\underline{a})\bar\theta_\tau\left(\frac{t}{2},y\right)dy\\ &+\frac{\underline{a}}{a(x)}\int_0^{\frac{t}{2}}\int_R G_{yyt}\cdot(a(y)-\underline{a})\bar\theta(\tau,y)dyd\tau\\ &+\frac{1}{a(x)}\int_{\frac{t}{2}}^{t}\int_R G\cdot(a(y)-\underline{a})\bar\theta_{\tau\tau}(\tau,y)dyd\tau.\end{aligned} \tag{3.43}$$

In a similar fashion to the previous estimates, we have

$$\sup_R |(\bar\theta-\bar\theta_0)_t(t,x)| \le Ct^{-\frac{7}{4}}. \tag{3.44}$$

Differentiating (3.43) in t, we also obtain

$$\sup_R |(\bar\theta-\bar\theta_0)_{tt}(t,x)| \le Ct^{-\frac{11}{4}}. \tag{3.45}$$

Here, we go back to (3.37). By virtue of (3.42) and (3.44),

$$\begin{cases}\sup_R|\bar\theta(t,x)| \le \sup_R|\bar\theta_0(t,x)| + Ct^{-\frac{3}{4}} \le Ct^{-\frac{1}{2}},\\ \sup_R|\bar\theta_t(t,x)| \le \sup_R|(\bar\theta_0)_t(t,x)| + Ct^{-\frac{7}{4}} \le Ct^{-\frac{3}{2}}.\end{cases} \tag{3.46}$$

Therefore, applying (3.46), instead of (3.33) and (3.34), to (3.37), we obtain

$$\sup_R |\bar\theta(t,x)-\bar\theta_0(t,x)| \le Ct^{-1}. \tag{3.47}$$

Similarly, we have

$$\sup_R |(\bar\theta-\bar\theta_0)_t(t,x)| \le Ct^{-2}. \tag{3.48}$$

However, this method is not applicable to $(\bar\theta-\bar\theta_0)_{tt}$ because we have $\bar\theta_{ttt}$ in the expression of $(\bar\theta-\bar\theta_0)_{tt}$.

The L^2-estimates to $\bar\theta-\bar\theta_0$ are also obtained by applying the Hausdorff–Young inequality. Thus we have the following lemma.

Lemma 3.2. *In addition to the assumption in Lemma* 3.1*, if* $(\theta_0, s-\underline{s}) \in L^1(R)\times L^1(R)$*, then it holds that as* $t\to\infty$*,*

$$\begin{cases}\|((\bar\theta-\bar\theta_0),(\bar\theta-\bar\theta_0)_t,(\bar\theta-\bar\theta_0)_{tt})(t,\cdot)\|_{L^\infty} = O(t^{-1},t^{-2},t^{-\frac{11}{4}}),\\ \|((\bar\theta-\bar\theta_0),(\bar\theta-\bar\theta_0)_t,(\bar\theta-\bar\theta_0)_{tt})(t,\cdot)\|_{L^2} = O(t^{-\frac{3}{4}},t^{-\frac{7}{4}},t^{-\frac{5}{2}}).\end{cases} \tag{3.49}$$

Third step. We now turn to (3.2). The perturbation $\Theta := \theta-\bar\theta$ satisfies

$$\begin{cases}\Theta_t = (a(x)\Theta)_{xx} + \Phi(\theta,x)_{xx},\\ \Theta|_{t=0} = 0,\end{cases} \tag{3.50}$$

where

$$\Phi(\theta, x) = -\{p(\theta + \bar{v}, s) - p(\bar{v}, s) - p_v(\bar{v}, s)\theta\}. \tag{3.51}$$

Similar to (3.37), we have the expression

$$\begin{aligned}
\Theta(t, x) &= \frac{1}{a(x)} \int_R G\left(\frac{t}{2}, x - y\right)(a(y) - \underline{a})\Theta\left(\frac{t}{2}, y\right) dy \\
&\quad + \frac{1}{a(x)} \int_0^{\frac{t}{2}} \int_R G_t \cdot (a(y) - \underline{a})\Theta(\tau, y) dy d\tau \\
&\quad + \frac{1}{a(x)} \int_{\frac{t}{2}}^{t} \int_R G \cdot (a(y) - \underline{a})\Theta_\tau(\tau, y) dy d\tau \\
&\quad + \frac{\underline{a}}{a(x)} \int_0^{\frac{t}{2}} \int_R G_{yy} \cdot \Phi dy d\tau + \frac{1}{a(x)} \int_{\frac{t}{2}}^{t} \int_R G_t \cdot \Phi(\tau, y) dy d\tau \\
&\equiv II_1 + II_2 + II_3 + II_4 + II_5.
\end{aligned} \tag{3.52}$$

We estimate each term in (3.52). By virtue of (3.22), (3.23), and (3.46),

$$\begin{aligned}
\|\Theta(t)\|_{L^\infty} &\leq \|\theta(t)\|_{L^\infty} + \|\bar{\theta}(t)\|_{L^\infty} \\
&\leq C(1 + t)^{-\frac{1}{4}}
\end{aligned} \tag{3.53}$$

and

$$\begin{aligned}
\|\Theta_t(t)\|_{L^\infty} &\leq \|\theta_t(t)\|_{L^\infty} + \|\bar{\theta}_t(t)\|_{L^\infty} \\
&\leq C(1 + t)^{-\frac{5}{4}}.
\end{aligned} \tag{3.54}$$

Hence II_1–II_3 are easily estimated as

$$|II_1, II_2, II_3| \leq Ct^{-\frac{3}{4}}. \tag{3.55}$$

Estimates of II_4 and II_5 are as follows:

$$\begin{aligned}
|II_4| &\leq C \int_0^{\frac{t}{2}} \sup_R |\Phi| \|G_{yy}(t - \tau)\|_{L^1} d\tau \\
&\leq C \int_0^{\frac{t}{2}} \sup_R |\theta|^2 (t - \tau)^{-1} d\tau \\
&\leq Ct^{-\frac{1}{2}}
\end{aligned} \tag{3.56}$$

and

$$
\begin{aligned}
|II_5| &\leq C \left| -\left[\int_R G\Phi(\tau,y)dy\right]_{\frac{t}{2}}^{t} + \int_{\frac{t}{2}}^{t}\int_R G\Phi_\tau(\tau,y)dyd\tau \right| \\
&\leq C\left\{ |\Phi(\theta,x)| + \left|\int_R G\left(\frac{t}{2},x-y\right)\Phi dy\right| \right. \\
&\qquad\qquad \left. + \int_{\frac{t}{2}}^{t} \sup_R |\Phi_\tau(\tau,y)| \|G(t-\tau)\|_{L^1} d\tau \right\} \\
&\leq C\left\{ \sup_R |\theta|^2 + \int_{\frac{t}{2}}^{t} \sup_R|\theta| \sup_R |\theta_t| d\tau \right\} \\
&\leq Ct^{-\frac{1}{2}}.
\end{aligned} \tag{3.57}
$$

Therefore, we obtain

$$
\|\Theta(t)\|_{L^\infty} \leq C(1+t)^{-\frac{1}{2}}, \tag{3.58}
$$

and hence

$$
\|\theta(t)\|_{L^\infty} \leq \|\Theta(t)\|_{L^\infty} + \|\bar{\theta}(t)\|_{L^\infty} \leq C(1+t)^{-\frac{1}{2}}. \tag{3.59}
$$

Applying (3.54) and (3.58), instead of (3.53), to (3.52) again, we have

$$
\|\Theta(t)\|_{L^\infty} \leq C(1+t)^{-\frac{3}{4}}. \tag{3.60}
$$

If we obtain faster decay of $\|\Theta_t(t)\|_{L^\infty}$ than (3.54), then $\|\Theta(t)\|_{L^\infty}$ will decay faster than (3.60). Hence we next estimate $\|\Theta_t(t)\|_{L^\infty}$ using the explicit formula. Since Θ_t satisfies

$$
\begin{cases}
(\Theta_t)_t = (\underline{a}\Theta_t)_{xx} + \{(a(x)-\underline{a})\Theta_t\}_{xx} + (\Phi(\theta,x)_t)_{xx}, \\
\Theta_t|_{t=0} = \Phi(\theta_0,x)_{xx},
\end{cases} \tag{3.61}
$$

we also have the expression, similar to (3.52),

$$
\begin{aligned}
\Theta_t(t,x) &= \frac{\underline{a}}{a(x)}\int_R G\left(\frac{t}{2},x-y\right)(a(y)-\underline{a})\Theta_t\left(\frac{t}{2},y\right)dy \\
&\quad + \frac{1}{a(x)}\int_0^{\frac{t}{2}}\int_R G_t\cdot(a(y)-\underline{a})\Theta_t(\tau)dyd\tau + \frac{1}{a(x)}\int_{\frac{t}{2}}^{t}\int_R G\cdot(a(y)-\underline{a})\Theta_{tt}(\tau)dyd\tau \\
&\quad + \frac{1}{a(x)}\left\{\int_R G_t\left(\frac{t}{2},x-y\right)\Phi\left(\frac{t}{2}\right)dy + \int_0^{\frac{t}{2}}\int_R G_{tt}\cdot\Phi dyd\tau\right\} \\
&\quad + \frac{1}{a(x)}\left\{\int_{\frac{t}{2}}^{t}\int_R G\cdot\Phi_{tt}(\tau,y)dyd\tau + \int_R G\left(\frac{t}{2},x-y\right)\Phi_t dy - \Phi_t(\theta,x)\right\}. \\
&\equiv III_1 + III_2 + III_3 + III_4 + III_5.
\end{aligned} \tag{3.62}
$$

Applying (3.54) and (3.59), we obtain

$$
\begin{aligned}
|III_1| &\le C\sup_R\left|G\left(\frac{t}{2}\right)\right|\sup_R|\Theta_t|\|a(\cdot)-\underline{a}\|_{L^1}\\
&\le Ct^{-\frac{7}{4}}, \qquad (3.63)
\end{aligned}
$$

$$
\begin{aligned}
|III_2| &\le C\|a(\cdot)-\underline{a}\|_{L^1}\int_0^{\frac{t}{2}}\sup_R|G_t(t-\tau)|\sup_R|\Theta_t(\tau,y)|d\tau\\
&\le C\int_0^{\frac{t}{2}}(t-\tau)^{-\frac{3}{2}}(1+\tau)^{-\frac{5}{4}}d\tau \qquad (3.64)\\
&\le Ct^{-\frac{7}{4}},
\end{aligned}
$$

and

(3.65)

$$
\begin{aligned}
|III_4| &\le C\left\{\sup_R\left|G_t\left(\frac{t}{2}\right)\right|\sup_R\left|\Phi\left(\frac{t}{2},y\right)\right|+\int_0^{\frac{t}{2}}\|G_{tt}(t-\tau)\|_{L^1}\sup_R|\Phi(\tau,y)|d\tau\right\}\\
&\le C\left\{(1+t)^{-1-1}+\int_0^{\frac{t}{2}}(t-\tau)^{-2}(1+\tau)^{-1}d\tau\right\}\\
&\le C(1+t)^{-2}\log(2+t).
\end{aligned}
$$

Since

$$
\|\Theta_{tt}(t)\|_{L^\infty}\le\|\theta_{tt}(t)\|_{L^\infty}+\|\bar{\theta}_{tt}(t)\|_{L^\infty}\le C(1+t)^{-\frac{9}{4}}
$$

by (3.26), (3.49), and $\|(\bar{\theta}_0)_{tt}(t)\|_{L^\infty}\le C(1+t)^{-\frac{5}{2}}$, we have

$$
\begin{aligned}
|III_3| &\le C\|a(\cdot)-\underline{a}\|_{L^1}\int_{\frac{t}{2}}^t\sup_R|G(t-\tau)|\sup_R|\Theta_{tt}(\tau,y)|d\tau\\
&\le C\int_{\frac{t}{2}}^t(t-\tau)^{-\frac{1}{2}}(1+\tau)^{-\frac{9}{4}}d\tau\le Ct^{-\frac{7}{4}} \qquad (3.66)
\end{aligned}
$$

and

$$
\begin{aligned}
|III_5| &\le C\bigg\{\int_{\frac{t}{2}}^t\|G(t-\tau)\|_{L^1}\sup_R|\Phi_{tt}(\tau)|d\tau+\left\|G\left(\frac{t}{2}\right)\right\|_{L^1}\sup_R\left|\Phi_t\left(\frac{t}{2}\right)\right|\\
&\qquad+\sup_R|\Phi_t(t,y)|\bigg\}\\
&\le C\bigg\{\int_{\frac{t}{2}}^t(\sup_R|\theta_t(\tau)|^2+\sup_R|\theta(\tau)|\sup_R|\theta_{tt}(\tau)|)d\tau \qquad (3.67)\\
&\qquad+\sup_R\left|\theta\left(\frac{t}{2}\right)\right|\sup_R\left|\theta_t\left(\frac{t}{2}\right)\right|+\sup_R|\theta(t)|\sup_R|\theta_t(t)|\bigg\}\\
&\le C\left\{\int_{\frac{t}{2}}^t(1+\tau)^{-\frac{5}{2}}d\tau+(1+t)^{-\frac{7}{4}}\right\}\le Ct^{-\frac{3}{2}}.
\end{aligned}
$$

Here we have used $|\partial_t \Phi(\theta, x)| \le C|\theta||\theta_t|, |\partial_{tt}\Phi(\theta, x)| \le C(|\theta_t|^2 + |\theta||\theta_{tt}|)$. Combining (3.62)–(3.67) we have

$$\|\Theta_t(t)\|_{L^\infty} \le Ct^{-\frac{3}{2}}, \tag{3.68}$$

and hence

$$\|\theta_t(t)\|_{L^\infty} \le \|\Theta_t(t)\|_{L^\infty} + \|\bar{\theta}_t(t)\|_{L^\infty} \le Ct^{-\frac{3}{2}}. \tag{3.69}$$

Applying (3.68), instead of (3.54), and (3.69) to (3.62) again, we have

$$\|\Theta_t(t)\|_{L^\infty} \le C(1+t)^{-\frac{7}{4}}. \tag{3.70}$$

We now go back to the estimate of $\|\Theta(t)\|_{L^\infty}$. Applying (3.60) and (3.70) to (3.52), we obtain the sharper estimate

$$\|\Theta(t)\|_{L^\infty} \le C(1+t)^{-1}\log(2+t). \tag{3.71}$$

By differentiating $(3.61)_1$ with respect to t, we have the explicit formula of Θ_{tt}, similar to (3.62). Estimating all terms, we obtain

$$\|\Theta_{tt}(t)\|_{L^\infty} \le C(1+t)^{-\frac{11}{4}}, \tag{3.72}$$

the details of which are omitted. If we apply (3.72) to (3.62), then we get

$$\|\Theta_t(t)\|_{L^\infty} \le C(1+t)^{-2}\log(2+t). \tag{3.73}$$

The L^2-estimates of Θ are also obtained by the Hausdorff–Young inequality:

$$\begin{cases} \|\Theta(t)\|_{L^2} \le C(1+t)^{-\frac{3}{4}}, \\ \|\Theta_t(t)\|_{L^2} \le C(1+t)^{-\frac{7}{4}}, \\ \|\Theta_{tt}(t)\|_{L^2} \le C(1+t)^{-\frac{5}{2}}. \end{cases} \tag{3.74}$$

Once more, applying (3.74) to (3.52) and (3.62), we obtain

$$\begin{cases} \|\Theta(t)\|_{L^\infty} \le C(1+t)^{-1}, \\ \|\Theta_t(t)\|_{L^\infty} \le C(1+t)^{-2}. \end{cases} \tag{3.75}$$

Thus we obtain the following lemma.

LEMMA 3.3. *In addition to the assumptions in Lemma* 3.1, *if* $(\theta_0, s - \underline{s}) \in L^1(R) \times L^1(R)$, *then it holds that as* $t \to \infty$,

$$\begin{cases} \|(\Theta, \Theta_t, \Theta_{tt})(t, \cdot)\|_{L^\infty} = O(t^{-1}, t^{-2}, t^{-\frac{11}{4}}), \\ \\ \|(\Theta, \Theta_t, \Theta_{tt})(t, \cdot)\|_{L^2} = O(t^{-\frac{3}{4}}, t^{-\frac{7}{4}}, t^{-\frac{5}{2}}). \end{cases} \tag{3.76}$$

Combining Lemmas 3.2 and 3.3, we conclude the proof of Proposition 3.2. □

Proof of Proposition 3.3. Rewrite (3.6) as

$$q_t - \underline{a}q_{xx} = -\{p_v(\theta + \bar{v}, s_0) + \underline{a}\}q_{xx} \tag{3.77}$$

to have the expression

(3.78)

$$q(t,x)=\int_R G(t,x-y)q_0(y)dy-\int_0^t\int_R G(t-\tau,x-y)\{p_v(\theta+\bar{v},s_0)+\underline{a}\}q_{yy}dyd\tau.$$

Here we put $q(0,x)=q_0(x)$ and $\underline{a}=-p_v(\underline{v},\underline{s})$. Integration by parts yields

(3.79)

$$\begin{aligned}&-\int_0^{\frac{t}{2}}\int_R G(t-\tau,x-y)\{p_v(\theta+\bar{v},s_0)+\underline{a}\}q_{yy}dyd\tau\\&\quad=\int_0^{\frac{t}{2}}\int_R G(t-\tau,x-y)(q_\tau-\underline{a}\theta_\tau)dyd\tau\\&\quad=\left\{\left[\int_R G(t-\tau,x-y)(q-\underline{a}\theta)dy\right]_0^{\frac{t}{2}}-\int_0^{\frac{t}{2}}\int_R G_\tau(t-\tau,x-y)(q-\underline{a}\theta)dyd\tau\right\}.\end{aligned}$$

Thus we obtain

(3.80)

$$\begin{aligned}(q-\bar{q}_0)(t,x)=&\int_R G\cdot(q-\underline{a}\theta)dy\Big|_{\tau=\frac{t}{2}}-\int_0^{\frac{t}{2}}\int_R G_\tau(t-\tau,x-y)(q-\underline{a}\theta)dyd\tau\\&-\int_{\frac{t}{2}}^t\int_R G(t-\tau,x-y)\{p_v(\theta+\bar{v},s_0)+\underline{a}\}q_{yy}dyd\tau\end{aligned}$$

and

(3.81)

$$\begin{aligned}(q-\bar{q}_0)_x(t,x)=&\int_R G_x\cdot(q-\underline{a}\theta)dy\Big|_{\tau=\frac{t}{2}}-\int_0^{\frac{t}{2}}\int_R G_{x\tau}(t-\tau,x-y)(q-\underline{a}\theta)dyd\tau\\&-\int_{\frac{t}{2}}^t\int_R G_x(t-\tau,x-y)\{p_v(\theta+\bar{v},s_0)+\underline{a}\}q_{yy}dyd\tau.\end{aligned}$$

Here we have used $\bar{q}_0(t,x)=\underline{a}\int_R G(t,x-y)\theta_0(y)dy$. Dividing the final term of (3.81) as

$$-\left(\int_{\frac{t}{2}}^{t-1}+\int_{t-1}^t\right)\int_R G_x\cdot\{p_v(\theta+\bar{v},s_0)+\underline{a}\}q_{yy}dyd\tau=(i)+(ii),$$

we seek the L^∞-estimate of (i) and (ii). Noting that

$$\begin{cases}q-\underline{a}\theta=-\{p(\theta+\bar{v},s_0)-p(\bar{v},s_0)-p_v(\bar{v},s_0)\theta+(p_v(\bar{v},s_0)-p_v(\underline{v},\underline{s}))\theta\}\\ \qquad\quad=O(|\theta|^2+|a(x)-\underline{a}||\theta|),\\ p_v(\theta+\bar{v},s_0)+\underline{a}=p_v(\theta+\bar{v},s_0)-p_v(\bar{v},s_0)+p_v(\bar{v},s_0)-p_v(\underline{v},\underline{s})\\ \qquad\qquad\qquad\quad=O(|\theta|+|a(x)-\underline{a}|),\end{cases}$$

we obtain

$$\begin{aligned}
|(i)| &\leq C\int_{\frac{t}{2}}^{t-1}\int_R |G|(|\theta|+|a(y)-\underline{a}|)|\theta_\tau|dyd\tau\\
&\leq C\int_{\frac{t}{2}}^{t-1}\|G(t-\tau)\|_{L^\infty}\left(\|\theta(\tau)\|_{L^2}\|\theta_\tau(\tau)\|_{L^2}+\|a(y)-\underline{a}\|_{L^1}\|\theta_\tau(\tau)\|_{L^\infty}\right)d\tau\\
&\leq C\int_{\frac{t}{2}}^{t-1}(t-\tau)^{-1}(1+\tau)^{-\frac{3}{2}}d\tau \leq Ct^{-\frac{3}{2}}\log(2+t)
\end{aligned}$$

and

$$\begin{aligned}
|(ii)| &\leq C\int_{t-1}^{t}\|G(t-\tau)\|_{L^2}\|\theta_\tau(\tau)\|_{L^\infty}\left(\|\theta(\tau)\|_{L^2}+\|a(y)-\underline{a}\|_{L^2}\right)d\tau\\
&\leq C\int_{t-1}^{t}(t-\tau)^{-\frac{3}{4}}(1+\tau)^{-\frac{3}{2}}d\tau \leq Ct^{-\frac{3}{2}}.
\end{aligned}$$

The other terms are estimated in a similar fashion to Proposition 3.2. The details are omitted. □

4. Pointwise estimate by the approximate Green function. In this final section, we devote ourselves to the proof of Theorem 2. In an expression obtained by differentiating V of (2.13) in x, $\int_{\frac{t}{2}}^{t}\int_R G_x\cdot\{(p_v(\tilde{v},s)+\underline{a})V_y\}_y dyd\tau$ is rewritten as follows:

$$\begin{aligned}
&-\int_{\frac{t}{2}}^{t}\int_R G_x\cdot\{(p_v(\tilde{v},s)+\underline{a})V_y\}_y dyd\tau\\
&=-\int_{\frac{t}{2}}^{t}\int_R G_{yy}\cdot(p_v(\tilde{v},s)+\underline{a})V_y dyd\tau\\
&=\frac{1}{\underline{a}}\int_{\frac{t}{2}}^{t}\int_R G_\tau\cdot(p_v(\tilde{v},s)+\underline{a})V_y dyd\tau\\
&=\frac{1}{\underline{a}}\left\{\left[\int_R G\cdot(p_v(\tilde{v},s)+\underline{a})V_y dy\right]_{\frac{t}{2}}^{t}-\int_{\frac{t}{2}}^{t}\int_R G\cdot\{(p_v(\tilde{v},s)+\underline{a})V_y\}_\tau dyd\tau\right\}\\
&=\frac{(p_v(\tilde{v},s)+\underline{a})}{\underline{a}}V_x(t,x)-\frac{1}{\underline{a}}\int_R G\cdot(p_v(\tilde{v},s)+\underline{a})V_y dy\Big|_{\tau=\frac{t}{2}}\\
&\quad-\frac{1}{\underline{a}}\int_{\frac{t}{2}}^{t}\int_R G\cdot\{(p_v(\tilde{v},s)+\underline{a})V_y\}_\tau dyd\tau,
\end{aligned}$$

where G is defined in (2.6) and $\underline{a}=-p_v(\underline{v},\underline{s})$. Hence

$$
\begin{aligned}
V_x(t,x) &= \frac{\underline{a}}{-p_v(\tilde{v},s)}\int_R G_x(t,x-y)V_0(y)dy \\
&\quad -\frac{\underline{a}}{-p_v(\tilde{v},s)}\left(\int_0^{\frac{t}{2}}+\int_{\frac{t}{2}}^t\right)\int_R G_x\{F_y+V_{\tau\tau}\}dyd\tau \\
&\quad +\frac{\underline{a}}{-p_v(\tilde{v},s)}\left(\int_0^{\frac{t}{2}}+\int_{\frac{t}{2}}^t\right)\int_R G_x p(\tilde{v},s)_{y\tau}dyd\tau \\
&\quad -\frac{1}{-p_v(\tilde{v},s)}\int_R \left(p_v(\tilde{v},s)+\underline{a}\right)G\left(\frac{t}{2}\right)V_y\left(\frac{t}{2}\right)dy \\
&\quad -\frac{\underline{a}}{-p_v(\tilde{v},s)}\int_0^{\frac{t}{2}}\int_R \left(p_v(\tilde{v},s)+\underline{a}\right)G_{yy}V_y dyd\tau \\
&\quad -\frac{1}{-p_v(\tilde{v},s)}\int_{\frac{t}{2}}^t\int_R \left\{\left(p_v(\tilde{v},s)+\underline{a}\right)V_y\right\}_\tau Gdyd\tau \\
&= J_1+(J_{21}+J_{22})+(J_{31}+J_{32})+J_4+J_5+J_6.
\end{aligned}
\tag{4.1}
$$

By $V_0 \in L^1(R)$, it is easy to see

$$
|J_1| \le Ct^{-1}. \tag{4.2}
$$

To estimate all other terms, we use (2.5) in Proposition 2.1 and (2.11) in Theorem 1. Integration by parts in x yields

$$
\begin{aligned}
|J_{21}| &\le C\left|\int_0^{\frac{t}{2}}\left(\int_R G_{yy}Fdy+\int_R G_y z_\tau dy\right)d\tau\right| \\
&\le C\left[\int_0^{\frac{t}{2}}\int_R |G_{yy}||F|dyd\tau+\left|\left[\int_R G_y zdy\right]_0^{\frac{t}{2}}-\int_0^{\frac{t}{2}}\int_R G_{y\tau}zdyd\tau\right|\right] \\
&\le C\left[\int_0^{\frac{t}{2}}\sup_R|G_{yy}(t-\tau)|\|V_y(\tau)\|^2d\tau+\left\|G_y\left(\frac{t}{2}\right)\right\|\left\|z\left(\frac{t}{2}\right)\right\|\right. \\
&\qquad \left.+\sup_R|G_y(t)|\|z(0)\|_{L^1}+\int_0^{\frac{t}{2}}\|G_\tau(t-\tau)\|\|z_y(\tau)\|d\tau\right] \\
&\le C\left[t^{-\frac{3}{2}}\int_0^{\frac{t}{2}}\|V_y(\tau)\|^2d\tau+t^{-\frac{3}{4}-1}+t^{-1}+t^{-\frac{5}{4}}\int_0^{\frac{t}{2}}(1+\tau)^{-\frac{3}{2}}d\tau\right] \\
&\le Ct^{-1}
\end{aligned}
\tag{4.3}
$$

and

$$
\begin{aligned}
|J_{22}| &\le C\left|\int_{\frac{t}{2}}^t\int_R G_{yy}Fdxd\tau+\int_{\frac{t}{2}}^t\int_R G_x z_\tau dxd\tau\right| \\
&\le C\left|\frac{1}{\underline{a}}\left(\left[\int_R GFdx\right]_{\frac{t}{2}}^t-\int_{\frac{t}{2}}^t\int_R GF_\tau dxd\tau\right)+\int_{\frac{t}{2}}^t\int_R G_x z_\tau dxd\tau\right|
\end{aligned}
$$

$$
\begin{aligned}
&\le C\Bigg\{\sup_R|F(t)|+\sup_R\left|F\left(\frac{t}{2}\right)\right|\left\|G\left(\frac{t}{2}\right)\right\|_{L^1}\\
&\qquad+\int_{\frac{t}{2}}^{t}\int_R(|V_y|^2|\tilde{v}|+|V_y||z_y|)|G|dxd\tau+\int_{\frac{t}{2}}^{t}\|G_x(t-\tau)\|\|z_\tau(\tau)\|d\tau\Bigg\}\\
&\le C\left\{\sup_R|V_x(t)|^2+\int_{\frac{t}{2}}^{t}\sup_R|V_y|^2\sup_R|\tilde{v}|\|G(t-\tau)\|_{L^1}d\tau\right.\\
&\qquad\left.+\int_{\frac{t}{2}}^{t}\sup_R|V_y|\|z_y(\tau)\|\|G(t-\tau)\|d\tau+\int_{\frac{t}{2}}^{t}\|G_x(t-\tau)\|\|z_\tau(\tau)\|d\tau\right\}\\
&\le C\left\{t^{-\frac{3}{2}}+t^{-\frac{3}{2}-\frac{3}{2}+1}+t^{-\frac{3}{4}-\frac{3}{2}+\frac{3}{4}}+t^{-\frac{3}{2}+\frac{1}{4}}\right\}\le Ct^{-\frac{5}{4}}.
\end{aligned}
\tag{4.4}
$$

For J_3, we have

$$
\begin{aligned}
|J_{31}|&\le C\left|\int_0^{\frac{t}{2}}\int_R G_{yy}p(\tilde{v},s)_\tau dyd\tau\right|\\
&\le C\int_0^{\frac{t}{2}}\|G_{yy}(t-\tau)\|_{L^1}\sup_R|\tilde{v}_\tau(\tau)|d\tau\\
&\le C\int_0^{\frac{t}{2}}(t-\tau)^{-1}(1+\tau)^{-\frac{3}{2}}d\tau\le Ct^{-1}
\end{aligned}
\tag{4.5}
$$

and

$$
\begin{aligned}
|J_{32}|&\le C\left|\int_{\frac{t}{2}}^{t}\int_R G_{yy}p(\tilde{v},s)_\tau dyd\tau\right|\\
&\le C\left|\int_{\frac{t}{2}}^{t}\int_R G_\tau p(\tilde{v},s)_\tau dyd\tau\right|\\
&=C\left|\left[\int_R Gp(\tilde{v},s)_\tau dy\right]_{\frac{t}{2}}^{t}-\int_{\frac{t}{2}}^{t}\int_R Gp(\tilde{v},s)_{\tau\tau}dyd\tau\right|\\
&\le C\Bigg\{\sup_R|\tilde{v}_\tau(t)|+\sup_R\left|\tilde{v}_\tau\left(\frac{t}{2}\right)\right|\left\|G\left(\frac{t}{2}\right)\right\|_{L^1}\\
&\qquad+\int_{\frac{t}{2}}^{t}\left(\sup_R|\tilde{v}_\tau|^2+\sup_R|\tilde{v}_{\tau\tau}|\right)\|G(t-\tau)\|_{L^1}d\tau\Bigg\}\\
&\le C\left(t^{-\frac{3}{2}}+t^{-\frac{5}{2}+1}\right)\le Ct^{-\frac{3}{2}}.
\end{aligned}
\tag{4.6}
$$

For the last three terms, we have

$$
\begin{aligned}
|J_4|&\le C\left(\sup_R|V_y(t)|\|\tilde{v}(t,\cdot)-\bar{v}(\cdot)\|\left\|G\left(\frac{t}{2}\right)\right\|+\|s(\cdot)-\underline{s}\|_{L^1}\sup_R\left|G\left(\frac{t}{2}\right)\right|\sup_R|V_y(t)|\right)\\
&\le C\left(t^{-\frac{3}{4}-\frac{1}{4}-\frac{1}{4}}+t^{-\frac{1}{2}-\frac{3}{4}}\right)\le Ct^{-\frac{5}{4}},
\end{aligned}
\tag{4.7}
$$

$$
\begin{aligned}
|J_5| &\le C\left(\int_0^{\frac{t}{2}}\sup_R|V_y(\tau)|\|\tilde{v}(t,\cdot)-\bar{v}(\cdot)\|\|G_{yy}(t-\tau)\|d\tau\right.\\
&\qquad\qquad \left.+\|s(\cdot)-\underline{s}\|_{L^1}\int_0^{\frac{t}{2}}\sup_R|G_{yy}(t-\tau)|\sup_R|V_y(\tau)|d\tau\right)\\
&\le C\left(\int_0^{\frac{t}{2}}(1+\tau)^{-\frac{3}{4}}(1+\tau)^{-\frac{1}{4}}(t-\tau)^{-\frac{5}{4}}d\tau+\int_0^{\frac{t}{2}}(t-\tau)^{-\frac{3}{2}}(1+\tau)^{-\frac{3}{4}}d\tau\right)\\
&\le C\left\{t^{-\frac{5}{4}}\log(2+t)+t^{-\frac{3}{2}+\frac{1}{4}}\right\}\le Ct^{-\frac{5}{4}}\log(2+t),
\end{aligned}
\tag{4.8}
$$

and

$$
\begin{aligned}
|J_6| \le C&\left(\int_{\frac{t}{2}}^{t}\sup_R|z_y(\tau)|\|G(t-\tau)\|\|\tilde{v}(t,\cdot)-\bar{v}(\cdot)\|d\tau\right.\\
&\qquad+\|s(\cdot)-\underline{s}\|_{L^1}\int_{\frac{t}{2}}^{t}\sup_R|z_y(\tau)|\sup_R|G(t-\tau)|d\tau\\
&\qquad\left.+C\int_{\frac{t}{2}}^{t}\sup_R|\tilde{v}_t(\tau)|\sup_R|V_y(\tau)|\|G(t-\tau)\|_{L^1}d\tau\right)\\
\le C&\left(t^{-\frac{3}{2}-\frac{1}{4}+\frac{3}{4}}+t^{-\frac{3}{2}+\frac{1}{2}}+t^{-\frac{3}{2}-\frac{3}{4}+1}\right)\le Ct^{-1}.
\end{aligned}
\tag{4.9}
$$

Combining (4.2)–(4.9) we have the desired rate

$$
\sup_R|V_x(t,x)|\le Ct^{-1}. \tag{4.10}
$$

Next, we estimate $\sup_x|z(t,x)|$ in a similar way to that above. Differentiating (2.13) in t, we have

$$
\begin{aligned}
z(t,x)=&\int_R G_t(t,x-y)V_0(y)dy+\int_R G\cdot p(\tilde{v},s)_{y\tau}dy\Big|_{\tau=\frac{t}{2}}\\
&-\left\{\int_0^{\frac{t}{2}}\int_R G_\tau\cdot p(\tilde{v},s)_{y\tau}dyd\tau-\int_{\frac{t}{2}}^{t}\int_R G\cdot p(\tilde{v},s)_{y\tau\tau}dyd\tau\right\}\\
&-\int_R G\cdot\{F_y+V_{\tau\tau}\}dy\Big|_{\tau=\frac{t}{2}}\\
&+\left\{\int_0^{\frac{t}{2}}\int_R G_t\{F_y+V_{\tau\tau}\}dyd\tau+\int_{\frac{t}{2}}^{t}\int_R G\cdot\{F_{yt}+V_{\tau\tau\tau}\}dyd\tau\right\}\\
&-\int_R\left\{(p_v(\tilde{v},s)+\underline{a})\,G_y\right\}_y Vdy\Big|_{\tau=\frac{t}{2}}\\
&-\int_0^{\frac{t}{2}}\int_R\left\{(p_v(\tilde{v},s)+\underline{a})\,G_y\right\}_{yt}Vdyd\tau\\
&-\int_{\frac{t}{2}}^{t}\int_R\left\{(p_v(\tilde{v},s)+\underline{a})\,V_y\right\}_\tau G_ydyd\tau.
\end{aligned}
\tag{4.11}
$$

Here we have used the integration by parts to gather the derivatives of G in the part on the domain $\left[0,\frac{t}{2}\right]$ and, on the other hand, to gather those in the other part on the domain $\left[\frac{t}{2}, t\right]$. By almost the same calculations as the estimate of $\sup_R |V_x(t,x)|$, we have the desired rate

$$\sup_R |z(t,x)| \leq Ct^{-\frac{3}{2}}. \tag{4.12}$$

Applying the Hausdorff–Young inequality, we also have the L^2-estimate

$$\|(V_x, z)(t,\cdot)\|_{L^2} = O(t^{-\frac{3}{4}}, t^{-\frac{5}{4}}),$$

which completes the proof of Theorem 2. □

Remark 6. We can also get

$$\|V(t,\cdot)\|_{L^\infty} \leq Ct^{-\frac{1}{2}}, \quad \|V(t,\cdot)\|_{L^2} \leq Ct^{-\frac{1}{4}}.$$

REFERENCES

[1] T. Gallay and G. Raugel, *Scaling variables and asymptotic expansions in damped wave equations*, J. Differential Equations, 150 (1998), pp. 42–97.

[2] L. Hsiao and D. Serre, *Large-time behavior of solutions for the system of compressible adiabatic flow through porous media*, Chinese Ann. Math. Ser. B, 16B (1995), pp. 431–444.

[3] L. Hsiao and D. Serre, *Global existence of solutions for the system of compressible adiabatic flow through porous media*, SIAM J. Math. Anal., 27 (1996), pp. 70–77.

[4] L. Hsiao and T.-P. Liu, *Convergence to nonlinear diffusion waves for solutions of a system of hyperbolic conservation laws with damping*, Comm. Math. Phys., 143 (1992), pp. 599–605.

[5] L. Hsiao and T. Luo, *Nonlinear diffusive phenomena of solutions for the system of compressible adiabatic flow through porous media*, J. Differential Equations, 125 (1996), pp. 329–365.

[6] L. Hsiao and T. Luo, *Nonlinear diffusive phenomena of entropy weak solutions for a system of quasilinear hyperbolic conservation laws with damping*, Quart. Appl. Math., 56 (1998), pp. 173–198.

[7] O.A. Ladyzhenskaja, V.A. Solonnikov, and N.N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, AMS, Providence, RI, 1968.

[8] P. Marcati and R. Pan, *On the diffusive profiles for the system of compressible adiabatic flow through porous media*, SIAM J. Math. Anal, to appear.

[9] K. Nishihara, *Convergence rates to nonlinear diffusion waves for solutions of system of hyperbolic conservation laws with damping*, J. Differential Equations, 131 (1996), pp. 171–188.

[10] K. Nishihara, *Asymptotic behavior of solutions of quasilinear hyperbolic equations with linear damping*, J. Differential Equations, 137 (1997), pp. 384–395.

[11] K. Nishihara and T. Yang, *Boundary effect on asymptotic behavior of solutions to the p-system with linear damping*, J. Differential Equations, 156 (1999), pp. 439–458.

SIAM J. MATH. ANAL.
Vol. 33, No. 1, pp. 240–259

ON THE PLACEMENT OF AN OBSTACLE OR A WELL SO AS TO OPTIMIZE THE FUNDAMENTAL EIGENVALUE[*]

EVANS M. HARRELL II[†], PAWEL KRÖGER[‡], AND KAZUHIRO KURATA[§]

Abstract. We investigate how to place an obstacle B within a domain Ω in Euclidean space so as to maximize or minimize the principal Dirichlet eigenvalue for the Laplacian on $\Omega \setminus B$. The shape of B is fixed a priori (usually as a ball), and only its position varies. We establish that for a certain class of domains the minimizing B is in contact with $\partial\Omega$, while the maximizing B is in the interior, typically at the center (supposing that the domain is sufficiently symmetric for this statement to be meaningful). Under special circumstances we can characterize the optimizing configurations with multiple obstacles. Our method relies on the Hadamard perturbation formula and a moving plane analysis.

Similar facts are proved when the hard obstacle is replaced by a central nonnegative potential function supported in B, and we consider the Schrödinger operator with this potential. Complementary facts are proved when the obstacle is replaced by a central nonpositive potential function.

Key words. Schrödinger equation, optimal eigenvalue, Hadamard, obstacle

AMS subject classifications. 47A75, 81Q10, 47A55, 49N45

PII. S0036141099357574

1. Introduction.

In this article we study how to minimize or maximize the fundamental eigenvalue of the Laplacian or Schrödinger operator defined within a fixed, bounded, open domain Ω, with zero Dirichlet boundary conditions on the boundary. Inside this domain we shall place an obstacle or a well, the position of which is under our control, and our goal is to locate the optimal position of the piece under our control.

The obstacles we consider may be hard, by which we mean that zero Dirichlet conditions are additionally imposed on the boundary of some open subset B of Ω, or they may be soft, by which we mean that the operator we analyze is of the following form:

$$-\nabla^2 \; + \; \alpha \; \chi_B(\mathbf{x}), \tag{1.1}$$

where $\alpha > 0$ and χ_B is the indicator function of the region B. Loosely, a hard obstacle corresponds to $\alpha = +\infty$. The term "well" refers to the situation where the constant α in operator (1.1) is negative. These operators are defined in standard ways (e.g., [Da95]), and by our sign convention the fundamental eigenvalue with an obstacle is positive and denoted λ; in the case of a well, λ might be negative. We recall that λ is nondegenerate and has an eigenfunction u which does not change sign. As usual, we choose $u(\mathbf{x}) > 0$ and normalize it in L^2 on Ω (respectively, on $\Omega \setminus B$ in the case of a hard obstacle).

[*] Received by the editors June 6, 1999; accepted for publication (in revised form) October 26, 2000; published electronically May 16, 2001.
http://www.siam.org/journals/sima/33-1/35757.html

[†] School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160 (harrell@math.gatech.edu). This author's work was supported by the Centre National pour la Recherche Scientifique and by NSF grant DMS-9622730.

[‡] Departamento de Matemática, Universidad Técnica Federico Santa María, Valparaíso, Chile (pkroeger@mat.utfsm.cl).

[§] Department of Mathematics, Tokyo Metropolitan University, Minami-Ohsawa 1-1, Hachioji-shi, Tokyo 192-0397, Japan (kurata@comp.metro-u.ac.jp). This author's work was supported by Grant-in-Aid for Scientific Research (C) 09640208.

Of particular interest for the light they shed on the relationship of geometry to the fundamental eigenvalue are the following questions about the placement of the interior obstacle:

1. Is it true that the optimal placement of an obstacle so as to minimize λ is in contact with the boundary, while the optimal placement to maximize λ is in the interior?

2. Given an affirmative answer to question 1, can the optimal position be located precisely?

For the placement of wells, we pose the same questions, with our expectations regarding the minimizing and maximizing positions reversed. In this article we shall describe some circumstances when the first question can be answered affirmatively, and some more narrow circumstances when the second question can be answered.

These expectations are suggested by perturbative analysis. If either α or the size of B is small, then to leading order in perturbation theory the effect of adding a soft obstacle or well is asymptotic to

$$\alpha \int_B u^2 d^n x\,.$$

Thus (for example, considering the case with $\alpha > 0$) the strategy to minimize λ is to place the obstacle near the boundary, while the strategy to maximize λ is to place the obstacle in the interior, near the maximum of $u(x)$. The situation with a small hard obstacle is similar (using the estimates in [Fl95]), while the situation with a well is reversed.

In many of the situations in this article the inside region B will be a ball. It is clear that for many purposes it is only necessary for B to have certain reflection symmetries, but we have preferred to focus on the case where the statement of the result is simplest. (See section 4, Example 7.)

In 1995, E. B. Davies asked two of us (E.H. and P.K.) questions of this type, for a hard spherical obstacle within a sphere. We answered the questions privately, using methods like those of this article: The minimizing position of the interior sphere is at the boundary of Ω, while the maximizing position is at the center of the exterior Ω. Quite recently we have learned from Ashbaugh and Chatelain [As99] that in response to the same query from A. G. Ramm, they have answered it with similar methods.

Independently, one of us (K.K.) had been considering the problem of placing a positive potential with a specified integral within a region Ω, so as to minimize λ. This work appears in [CGIKO99].

In both these independent lines of investigation, the minimizing obstacles are in contact with the boundary. One aim here is to explore this phenomenon further.

We are not aware of other work on this problem, although there are some asymptotic estimates for small obstacles (especially [Fl95]), and some work on optimization of capacity on annular domains in [Fl93] and [Co94].

Our technique is to treat the motion of the obstacle or well as a perturbation, and estimate the perturbation with a reflection technique reminiscent of the classical method of moving planes [Al60], [Se71], [BeNi91], which, curiously, has heretofore not been used as much in spectral theory as in nonlinear analysis.

The first insight we use is that a translation of the obstacle or well can be regarded as a perturbation of a boundary. For a hard obstacle, the Hadamard boundary perturbation formula [Ha08], [GaSc53] applies. When specialized to the case of a translation, it reads simply as follows.

Proposition 1.1 (the Hadamard perturbation formula, special case). *Let B be an interior hard obstacle which can be moved rigidly a positive distance in the direction of a unit vector* $\mathbf{v}$*. The boundary of B is assumed piecewise smooth. Then*

$$\frac{\partial \lambda}{\partial \mathbf{v}} = \int_{\partial B} |\nabla u|^2 \, \mathbf{n} \cdot \mathbf{v} \, dS. \tag{1.2}$$

Here and throughout, $\mathbf{n}$ is the unit normal at the surface of the obstacle B. Our choice of orientation is outward with respect to B and hence inward with respect to Ω. We recall for later purposes that at a boundary with zero Dirichlet conditions, the gradient of u is parallel to the normal vector, provided the latter is defined.

When $\alpha > 0$ is finite, the derivative of the eigenvalue with respect to a translation of a soft obstacle is obtainable from Green's theorem. As for Hadamard's formula, the derivative is proportional to a certain surface integral. As a first step in deriving the formula for this derivative, we prepare an estimate for the eigenvalue $\lambda(\epsilon)$ and L^2-normalized eigenfunction $u(\epsilon)$ of a soft obstacle $B(\epsilon)$ obtained by shifting B by a distance ϵ.

Lemma 1.2. (a) *There exist constants $\epsilon_0 > 0$ and C, depending only on Ω, such that*

$$|\lambda(\epsilon) - \lambda| \leq C\epsilon$$

for every $0 < \epsilon < \epsilon_0$.

(b) $\|u(\epsilon) - u\|_{L^\infty(\Omega)} \to 0$ *as $\epsilon \to 0$.*

Proof. According to the min-max principle,

$$\lambda(\epsilon) \leq \int_\Omega |\nabla u|^2 \, dx + \alpha \int_\Omega \chi_{B(\epsilon)} u^2 \, dx = \lambda + \alpha \int_\Omega (\chi_{B(\epsilon)} - \chi_B) u^2 \, dx.$$

Recall that u and $u(\epsilon) \in L^\infty(\Omega)$ with $\|u(\epsilon)\|_\infty \leq C_1$, with C_1 depending only on Ω (see, e.g., [GiTr83, Theorem 8.15]). Since also

$$\int_\Omega |(\chi_{B(\epsilon)} - \chi_B)| \, dx \leq C_2 \epsilon,$$

we obtain

$$\lambda(\epsilon) - \lambda \leq C_3 \epsilon.$$

By an analogous argument, $\lambda - \lambda(\epsilon) \leq C_4 \epsilon$, completing the proof of (a).

Next, it is easy to see that $\{u(\epsilon)\}_{0<\epsilon<\epsilon_0}$ is bounded in $H_0^1(\Omega)$, so there exists a subsequence of $\{u(\epsilon)\}_{0<\epsilon<\epsilon_0}$ which converges weakly in $H_0^1(\Omega)$. Actually, $u(\epsilon)$ must converge to u weakly in $H_0^1(\Omega)$ and strongly in $L^2(\Omega)$, since the weak limit of any subsequence of $\{u(\epsilon)\}_{0<\epsilon<\epsilon_0}$ is the first eigenfunction u associated to λ and hence unique. Let $w(\epsilon) := u(\epsilon) - u$. Then $w(\epsilon)$ satisfies

$$-\Delta w(\epsilon) + (\alpha \chi_{B(\epsilon)} - \lambda) w(\epsilon) = (\lambda(\epsilon) - \lambda) u(\epsilon) - \alpha(\chi_{B(\epsilon)} - \chi_B) u.$$

Applying the uniform L^∞ estimate from [GiTr83, Theorem 8.15] to $w(\epsilon)$, we have

$$\|w(\epsilon)\|_{L^\infty(\Omega)} \leq \|w(\epsilon)\|_{L^2(\Omega)} + C_5 |\lambda(\epsilon) - \lambda| + C_6 \|\chi_{B(\epsilon)} - \chi_B\|_{L^1(\Omega)}.$$

This yields the desired estimate. □

We shall need the following elementary formula.

LEMMA 1.3. *Suppose $\zeta \in C^1(\Omega)$. Then*

$$\frac{1}{\epsilon}\left(\int_{B(\epsilon)} \zeta^2\, dx - \int_B \zeta^2\, dx\right) \to \int_{\partial B} \zeta^2 (\mathbf{v}\cdot\mathbf{n})\, dS$$

as $\epsilon \to 0$.

Proof. Since $\int_{B(\epsilon)} \zeta^2\, dx = \int_B \zeta^2(\mathbf{y}+\epsilon\mathbf{v})\, dy$, we get

$$\begin{aligned}\frac{1}{\epsilon}\left(\int_{B(\epsilon)} \zeta^2\, dx - \int_B \zeta^2\, dx\right) &= \int_{B(\epsilon)}\left(\frac{\zeta(\mathbf{y}+\epsilon\mathbf{v})-\zeta(\mathbf{y})}{\epsilon}\right)(\zeta(\mathbf{y}+\epsilon(\mathbf{v}))+\zeta(\mathbf{y}))dy\\ &\to 2\int_B \frac{\partial\zeta}{\partial\mathbf{v}}\zeta dy = \int_B \nabla\cdot\ \left(\zeta^2\mathbf{v}\right) dy.\end{aligned}$$

The divergence theorem implies the desired result. □

PROPOSITION 1.4. *Consider the case of a soft obstacle or a well* (1.1), *where B is assumed to have a piecewise smooth boundary. Suppose that B can be moved rigidly a positive distance in the direction of a unit vector* $\mathbf{v}$. *Then*

$$\frac{\partial\lambda}{\partial\mathbf{v}} \;=\; \alpha\int_{\partial B} |u|^2\mathbf{n}\cdot\ \mathbf{v}\ dS. \tag{1.3}$$

Proof. We denote by $\lambda(\epsilon)$ and λ the fundamental eigenvalues of the operators $-\Delta+\alpha\chi_{B(\epsilon)}$ and $-\Delta+\alpha\chi_B$, respectively. Here $B(\epsilon)=\{\mathbf{y}\in\mathbf{R}^n;\mathbf{y}=\mathbf{x}+\epsilon\mathbf{v}\}$ for small $\epsilon>0$. We choose $0<\epsilon<\epsilon_0$, where $\epsilon_0>0$ is sufficiently small so that $B(\epsilon)\subset\Omega$. We also denote by $u(\epsilon)$ and u the L^2-normalized eigenfunctions associated with $\lambda(\epsilon)$ and λ, respectively. Thus $u(\epsilon)$ and u satisfy

$$\int_\Omega \nabla u(\epsilon)\cdot\nabla\phi + \alpha\chi_{B(\epsilon)}u(\epsilon)\phi\, dx = \lambda(\epsilon)\int_\Omega u(\epsilon)\phi\, dx$$

and

$$\int_\Omega \nabla u\cdot\nabla\psi + \alpha\chi_B u\psi\, dx = \lambda\int_\Omega u\psi\, dx$$

for every $\phi,\psi\in H_0^1(\Omega)$. Substituting $\phi=u$ and $\psi=u(\epsilon)$, we have

$$(\lambda(\epsilon)-\lambda)\int_\Omega u(\epsilon)u\, dx = \alpha\int_\Omega(\chi_{B(\epsilon)}-\chi_B)u(\epsilon)u\, dx.$$

By Lemma 1.2(b), we have

$$\left|\int_\Omega u(\epsilon)u\, dx - \int_\Omega u^2\, dx\right| = o(1)$$

and

$$\left|\int_\Omega(\chi_{B(\epsilon)}-\chi_B)(u(\epsilon)u-u^2)\, dx\right| = o(\epsilon).$$

Hence it follows from Lemma 1.2(a) that

$$\frac{(\lambda(\epsilon)-\lambda)}{\epsilon} = \frac{\alpha}{\epsilon}\int_\Omega(\chi_{B(\epsilon)}-\chi_B)u^2\, dx + o(1).$$

Since $u \in C^{1,\beta}(\Omega)$ for $0 < \beta < 1$ (see, e.g., [GiTr83]), Lemma 1.3 yields

$$\frac{d\lambda}{d\mathbf{v}} = \frac{d\lambda(\epsilon)}{d\epsilon}|_{\epsilon=0} = \alpha \int_{\partial B} u^2(\mathbf{v}\cdot\mathbf{n})\,dS. \qquad \square$$

Since the possible centers of the obstacle or well form a compact subset of Ω, it is immediate from Lemmas 1.2 and 1.3 that under the assumptions of this article, *the maximizing and minimizing positions of B exist.*

2. The technique of domain reflection. If a domain has a certain reflection property with respect to an axis of symmetry of the obstacle, then we shall be able to identify the sign of the directional derivative of the fundamental eigenvalue with respect to the position of an obstacle or well. Roughly speaking, when this property holds, we shall show that the eigenvalue increases as the obstacle moves away from a nearby portion of the boundary of Ω.

To avoid complications, we henceforth assume that the set B is convex as well as piecewise smooth. We also require that it be reflection-symmetric about some hyperplane (or plane, or line) P of dimension $n-1$. When we consider specific examples, B will often be a ball.

Definition. *Let P be a hyperplane of dimension $n-1$ which intersects Ω. For any connected set S which does not intersect P, we let S^P denote its reflection through P. The domain Ω is said to have the* interior reflection property *with respect to P if there is a connected component Ω_s of $\Omega \setminus P$ such that Ω_s^P is a proper subset of the other connected component Ω_b of $\Omega \setminus P$. Any such P will be called a* hyperplane of interior reflection *for Ω. Moreover, Ω_s will be called the* small side *of Ω (and Ω_b will be called the* big side*).*

The following theorem states formally that when this property holds, the eigenvalue is strictly increasing as a symmetric obstacle is moved away from the small side.

Theorem 2.1. *Assume that Ω has the interior reflection property with respect to a hyperplane P about which the set B is reflection-symmetric. Suppose that B is translated in the direction of a unit vector $\mathbf{v}$ perpendicular to P and pointing from the small side to the big side.*

Then, in the case of a hard or soft obstacle,

$$\frac{d\lambda}{d\mathbf{v}} \; > \; 0.$$

In the case of a well,

$$\frac{d\lambda}{d\mathbf{v}} \; < \; 0.$$

Remark. Actually, the soft obstacle or well here could be any reflection-symmetric function supported within the closure of B, not just its indicator function.

Proof. There are three cases to consider, that of a hard obstacle, a soft obstacle, and a well. We consider the hard obstacle last.

For the other two cases, we claim that for any point $\mathbf{x}$ of ∂B which is on the small side of Ω, $u(\mathbf{x}) < u(\mathbf{x}^P)$. The theorem will then follow in these cases from (1.3).

To establish the claim, we consider $w(\mathbf{x}) := u(\mathbf{x}) - u\left(\mathbf{x}^P\right)$ on the small side Ω_s. On the interior of this region,

$$\left(-\nabla^2 \; + \; \alpha\chi_B\right) w \; = \; \lambda\, w,$$

while on its boundary, $w(\mathbf{x}) \leq 0$. Observing that w is strictly negative on part of that boundary and that λ is less than the fundamental Dirichlet eigenvalue of $-\nabla^2 + \alpha\chi_B$, we conclude from the maximum principle [PrWe84] that $w(\mathbf{x}) < 0$ in the interior of this region, and hence that $u(\mathbf{x}) < u(\mathbf{x}^P)$ for $\mathbf{x}$ in ∂B on the small side of Ω.

This establishes the claim except for the case of a hard obstacle, where we use the Hadamard formula (1.2) in place of (1.3). This time we consider the function $w(\mathbf{x}) := u(\mathbf{x}) - u(\mathbf{x}^P)$ on the small side Ω_s but excluding B. Just as before, the maximum principle tells us that $w(\mathbf{x}) < 0$ on the interior of this region. To finish the proof in this case, we appeal to the boundary-point lemma of [Se71, p. 308], according to which, at every smooth point of the part of ∂B on the small side, either the normal derivative of $w(\mathbf{x})$ is strictly positive or else the second derivative of $w(\mathbf{x})$ in this direction is strictly positive. However, the latter possibility is excluded because it contradicts the eigenvalue equation (since the Laplacian of w is negative while all second derivatives in tangential directions at the boundary are 0). Hence $|\nabla u(\mathbf{x})| < |\nabla u(\mathbf{x}^P)|$ for $\mathbf{x}$ in ∂B on the small side of Ω. The theorem then follows from (1.2). □

In the final section of this article we consider many examples where it can be shown that either the domain Ω or a suitable related domain Ω' contains a dense subset of points which lie on a hyperplane of interior reflection. In preparation for that we state here an obvious corollary of Theorem 2.1.

Corollary 2.2. *Let $\mathbf{x} \in \Omega$ denote the center of a spherical obstacle B. At any maximizing $\mathbf{x}$,*

(a) *Ω has no hyperplane P of interior reflection containing $\mathbf{x}$.*

Moreover, at any minimizing $\mathbf{x}$, either statement (a) *above is true, or else*

(b) *∂B intersects the small side of $\partial\Omega$.*

Convexity ensures that Ω enjoys the interior reflection property with respect to some secant plane passing through any point sufficiently close to the boundary. This immediately implies the following.

Corollary 2.3. *Assume that Ω is convex and that B is a ball of radius ρ. There exists $R_0 > 0$ depending on Ω such that if $\rho < R_0$, then there are neighborhoods $N_{1,2}$ of the boundary, such that*

(a) *the maximizing* (*resp., minimizing*) *obstacle* (*resp., well*) *for λ lies outside N_1; and*

(b) *any obstacle* (*resp., well*) *which minimizes* (*resp., maximizes*) *λ subject to being located within N_2 must touch the boundary of Ω.*

In principle, given any convex Ω it is straightforward to identify neighborhoods $N_{1,2}$ explicitly. In the following section we consider some cases where $N_2 = \Omega$ and where the optimal positions can be determined exactly, sometimes even without convexity.

At the level of generality of Corollary 2.3 there is a "hole" in the interior of a convex Ω within which we can say little about the optimal placement of obstacles. With a reflection symmetry, however, the hole can be reduced to a slit, because together with convexity this implies that every point of Ω is either reflection-symmetric or else on a hyperplane of interior reflection.

Corollary 2.4. *Suppose Ω and the obstacle are as in* Corollary 2.3, *and in addition that Ω is symmetric with respect to reflection through a hyperplane H. Then at the minimizing position the obstacle is in contact with the boundary, while at the maximizing position its center is on H.*

Next we note an extension of Theorem 2.1 to the case of Schrödinger operators, which is useful for discussions of soft obstacles and of interest in its own right.

Theorem 2.5. *Consider the Schrödinger operator $\mathcal{H} := -\nabla^2 + V(\mathbf{x})$ on a domain Ω, where the potential $V(\mathbf{x})$ is a real-valued function in $L^\infty(\Omega)$ (or more generally satisfying conditions guaranteeing that the fundamental eigenvalue is discrete; see, e.g.,* [ReSi78]*). Assume that Ω has the interior reflection property with respect to a hyperplane P about which the set B is reflection-symmetric, and that on the small side Ω_s,*

$$V(\mathbf{x}) \geq V\left(\mathbf{x}^P\right) \text{ almost everywhere.}$$

Suppose that B is translated in the direction of a unit vector $\mathbf{v}$ perpendicular to P and pointing from the small side to the big side.

Then, in the case of a hard or soft obstacle (i.e., $-\nabla^2 + V(\mathbf{x}) + \alpha\, \chi_{\mathrm{B}}(\mathbf{x})$),

$$\frac{d\lambda}{d\mathbf{v}} > 0.$$

In the case of a well,

$$\frac{d\lambda}{d\mathbf{v}} < 0.$$

Proof. The proof is like that of Theorem 2.1. This time, however, $w(\mathbf{x}) := u(\mathbf{x}) - u\left(\mathbf{x}^P\right)$ is no longer a solution of the eigenvalue equation on Ω_s but is instead a subsolution, i.e.,

$$(H - \lambda)w(\mathbf{x}) = -\left(V\left(\mathbf{x}\right) - V\left(\mathbf{x}^P\right)\right) u\left(\mathbf{x}^P\right) \leq 0.$$

This, however, suffices for the maximum principle, by the following argument.

First we observe that $w(x) \leq 0$ on Ω_s as before. Indeed, if $U = \{w(x) > 0\}$ were nonempty, then the inequality in the proof of Theorem 2.5 would imply that $\lambda \geq \mu_1$, where μ_1 is the first eigenvalue of H on U. However, $U \subset \Omega_s \subset \Omega$ and λ is the first eigenvalue of H on Ω. As a consequence of the unique continuation theorem (e.g., [JeKe85]), we find that $\mu_1 > \lambda$, which is a contradiction.

Next, to conclude that $w(x) < 0$ in Ω_s, we appeal to the strong maximum principle. □

We close this section by observing that for sufficiently small α, any globally minimizing soft obstacle (resp., maximizing well) touches the boundary.

Theorem 2.6. *Suppose Ω is convex and that it contains a soft spherical obstacle, i.e., a potential $\alpha\, \chi_B(x)$, where B is a sufficiently small ball. Then there exists $\alpha_0 > 0$ such that for every α with $0 < \alpha < \alpha_0$, when B is at the position where it minimizes the first eigenvalue, it touches the boundary $\partial\Omega$.*

Proof. We let $B_\epsilon(\mathbf{w})$ denote the obstacle when centered at $\mathbf{w}$, and we assume that the radius $\epsilon > 0$ of the obstacle is sufficiently small.

First we claim that there exists a compact subset $G \subset \Omega$, independent of ϵ, such that if $\mathbf{w} \in G'$ and $\overline{B(\mathbf{w})} \subset \Omega$, then $B(\mathbf{w})$ cannot be an optimal obstacle. Here $G' := \Omega \setminus G$. This follows from Corollary 2.3. We observe that G may be chosen independently of ϵ for ϵ sufficiently small; the choice of G depends only on Ω.

Let ψ be the L^2-normalized first eigenfunction of the Dirichlet Laplacian on Ω, and let $[\Omega]^\delta = \{x \in \Omega; \psi(x) > \delta\}$ for $\delta > 0$. Then there exists a small $\delta > 0$ such that $G \subset\subset [\Omega]^\delta$. We fix a value of the radius $\epsilon > 0$ sufficiently small so that

$$\epsilon < \min(\mathrm{dist}(\partial G, \partial[\Omega]^\delta), (1/2)\mathrm{dist}(\partial\Omega, \partial[\Omega]^\delta)).$$

Next we assert that, as shown in [CGIKO99], there exists a small $\alpha_0 > 0$ such that if $\mathbf{w} \in G$, then $B_\epsilon(\mathbf{w})$ cannot be a minimizing obstacle for any $0 < \alpha < \alpha_0$, because

$$\|u(\epsilon, \mathbf{w}) - \psi\|_{L^\infty(\Omega)} \leq C\alpha$$

for some constant C which does not depend on $\mathbf{w}$.

Here $u(\epsilon, \mathbf{w})$ denotes the L^2-normalized first eigenfunction of $-\Delta + \alpha\chi_{B_\epsilon(\mathbf{w})}$. Note that if $\mathbf{w} \in G$, then $B_\epsilon(\mathbf{w}) \subset \{\psi(x) > \delta\}$, and we can find a ball $B_\epsilon(\mathbf{w}') \subset \{\psi(x) < \delta\}$. Hence, for sufficiently small α_0,

$$\int_{B_\epsilon(\mathbf{w})} u(\epsilon, \mathbf{w})^2 \, dx > \int_{B_\epsilon(\mathbf{w}')} u(\epsilon, \mathbf{w})^2 \, dx$$

for all $0 < \alpha < \alpha_0$. Together with the variational principle, this implies that $B_\epsilon(\mathbf{w})$ cannot be a minimizing obstacle for $\mathbf{w} \in G$ and for $0 < \alpha < \alpha_0$, completing the proof. □

3. Optimization at a vertex or corner. It is not difficult to show that an ellipse has the interior reflection property with respect to any secant line which is perpendicular to the boundary at one of its crossing points and which does not coincide with one of the axes. It thus follows fairly easily from Theorem 2.1 that if the radius of the ball B is sufficiently small so that it fits inside an elliptical domain Ω, then the minimizing ball touches the boundary. Actually, we can locate the minimizing position at the vertex of the ellipse, and the maximizing position at the center, for a class of domains generalizing the ellipse (see Theorem 3.2, below).

We shall show that this phenomenon, that minimizing obstacles are located at parts of the boundary where the curvature is maximized, also occurs in certain other situations. Unfortunately, we are not able to determine the degree of generality of this phenomenon.

We begin by extending Theorem 2.1 to the case where B moves along the boundary; to keep the statement simple, we restrict it to the case of spherical B.

PROPOSITION 3.1. *Let B be a ball which is tangent to the boundary of Ω, assumed of class C^2 in a neighborhood of the point of contact. Suppose furthermore that Ω has the interior reflection property with respect to a hyperplane P normal to the boundary at the point of contact. Then λ is strictly increasing as B is moved in contact with the boundary towards the big side.*

Sketch of the proof. The argument is by domain perturbation as for Theorem 2.1, with the further complication that as the domain B moves along a smooth boundary, it is not only translated and but also continuously rotated. For nonspherical domains, Propositions 1.1 and 1.4 would need to be modified with additional terms to reflect this. For spherical domains, however, the additional terms do not arise, and the formulae for the directional derivatives are as before. □

We next identify a class of roughly elliptical regions for which we can carry out a complete analysis of the maximizing and minimizing positions of an obstacle or well.

DEFINITION. *A* vertex *of a domain with boundary of class C^2 is a point on the boundary at which the curvature is locally maximal. Outward pointing corners of a piecewise C^2 boundary are also considered vertices.*

THEOREM 3.2. *Let Ω be a two-dimensional convex domain with the following properties:*

(a) *Ω is reflection symmetric with respect to both the x and y Cartesian axes.*

(b) *The boundary of* Ω *is of class* C^2 *for* $x, y \neq 0$.

(c) *In any quadrant of the plane, the curvature of the boundary of* Ω *is monotonic as a function of* x (*or equivalently of* y *or of the arclength* s).

Suppose that the obstacle (*resp., well*) B *is a disk.*

If the radius of B *is less than the radius of curvature at the vertex of* Ω, *then* λ *is minimized* (*resp., maximized*) *when* B *is in contact with a vertex, and maximized* (*resp., minimized*) *when the obstacle* (*resp., well*) *is at the origin.*

This theorem certainly generalizes to three-dimensional ellipsoidal domains Ω which are rotationally symmetric. We also remark that, as a special case, when both Ω and B are balls, λ is a strictly increasing function of the distance of B from the boundary until it reaches the center, where it is maximized. (This answers the query of Davies [Da95a].)

Theorem 3.2. is a direct corollary of the following.

Proposition 3.3. *A region as in Theorem* 3.2 *enjoys the interior reflection property with respect to any line normal to its boundary, except for the lines of symmetry* (*x and y axes*). *The small side of the normal line at a boundary point* P *is the side of increasing curvature of the boundary moving from* P.

Proof. We may and shall assume without loss of generality that the curvature of the boundary of Ω is strictly increasing as a function of y in the first quadrant of the plane (and hence strictly monotonic in any quadrant of the plane). We also orient the arclength s counterclockwise, so that s is an increasing function of y. We observe that Ω is convex and that in the first quadrant the distance from the origin to a point on $\partial\Omega$ increases with y. For simplicity we assume that $\partial\Omega$ is of class C^2 even at the possibly exceptional points ($x = 0$ or $y = 0$); it will be clear that this does not affect the result.

Let φ be the angle of the normal to the boundary as measured counterclockwise from the x-axis. Because of our assumptions, the angle φ can be used to parametrize the points of $\partial\Omega$; we henceforth do this, and use the notation $P(\varphi)$ for those points. Let L designate the normal line to $\partial\Omega$ at some specific $P(\alpha)$ in the first quadrant.

We claim first that if $\beta :=$ the angle $> \alpha$ where L intersects $\partial\Omega$, then $\beta < \pi + \alpha$. The statement that $\beta < \pi + \alpha$ is easily seen to be equivalent to the statement that L passes above the center of Ω, which in turn means that $\varphi < \theta$, where θ is the usual polar coordinate of $P(\alpha)$.

In order to show that the reflection of $\partial\Omega$ for $\alpha < \varphi_+ < \beta$ fits within Ω, we make some definitions. Let ℓ_+ be the distance from $P(\varphi_+)$ to L and let t be as the distance from $P(\alpha)$ to the point on L closest to $P(\varphi_+)$. At a given value of t, there is an analogous point $P(\varphi_-)$ on the other side of L, at some $\varphi_- < \alpha$; we define the distance from $P(\varphi_-)$ to L as ℓ_-.

We thus need to show that $\ell_+ \leq \ell_-$ for the same value of t. Hence we consider the maps $\varphi_\pm \to t$, for which

$$dt = \sin(\varphi - \alpha) ds = \frac{1}{\kappa(\varphi)} \left|\sin(\varphi - \alpha)\right| d\varphi.$$

Here, κ designates the curvature of the boundary at the point corresponding to φ. For $\alpha < \phi < \alpha + \frac{\pi}{2}$,

$$t(\varphi) - t(2\alpha - \varphi) = \int_\alpha^\varphi \left(\frac{1}{k(\phi)} - \frac{1}{k(2\alpha - \phi)}\right) |\sin(\phi - \alpha)| d\phi < 0 \qquad \text{for } \theta < \alpha + \frac{\pi}{2},$$

since the integrand is negative.

Furthermore, this difference is negative for all $\alpha < \varphi < \beta$ because the integral is antisymmetric about $\frac{\pi}{2} + \alpha$ and $\beta < \left(\frac{\pi}{2} + \alpha\right) + \left(\frac{\pi}{2} - \alpha\right) = \pi$. We conclude that $t(\varphi) > t(2\alpha - \varphi)$ ($2\alpha - \varphi$ is the angle reflected through α) for this range of φ, and therefore if we consider the two maps $\varphi \to t$, we see

$$|\varphi - \alpha|(t)_{\text{above } L} > |\varphi - \alpha|(t)_{\text{below } L}.$$

Also, both $t(\varphi)$ and $t(2\alpha - \varphi)$ are monotonic increasing functions of φ for $\varphi > \alpha$.

Since $\frac{d\ell_\pm}{ds} = \cot(\varphi - \alpha)$ and cot is a decreasing function for $0 < |\varphi - \alpha| < |\beta - \alpha| < \pi$, we conclude by integrating that

$$\ell_+ < \ell_- \qquad \text{for } 0 < t < t(\beta). \qquad \square$$

THEOREM 3.4. (a) *Suppose that Ω and B are as in Theorem* 3.2, *and that the radius of B is small enough for it to fit within Ω but larger than the radius of curvature at its vertex. Then λ is minimized* (*resp., maximized*) *when B is in as close as possible to a vertex and maximized* (*resp., minimized*) *when the obstacle* (*resp., well*) *is at the origin.*

(b) *Suppose that Ω is an equilateral polygon centered at the origin. Then λ is minimized* (*resp., maximized*) *when the ball is as close as possible to any vertex of Ω and maximized* (*resp., minimized*) *when the obstacle* (*resp., well*) *is at the origin.*

Proof. We discuss the case of an obstacle. As usual, the case of a well uses the same argument, with a reversal of "maximal" and "minimal."

Part (a) is an obvious variant of Theorem 3.2. It is necessary only to notice that if B is anywhere other than at the origin, the interior reflection holds with respect to the horizontal or vertical plane through the center of B. By Corollary 2.2, the only possibility for the maximizing position is the center, whereas at the minimizing position B is in contact with $\partial\Omega$.

To localize the minimizer more precisely, we observe that if the contact point is not a point of symmetry of $\partial\Omega$, and B is not obstructed from displacements to the small side of Ω, then Proposition 3.1 excludes the configuration as a candidate for a minimizer. It is moreover easy to see that any attainable point of symmetry other than a vertex is excluded from being a contact point of a minimizing obstacle. (It would in fact be a constrained maximizing position, given contact with the boundary.) The only remaining possibility is that described in statement (a).

The argument for part (b) is similar. From Theorem 2.1, using the symmetry of the polygon we see that if the center of B is anywhere other than at the center of Ω, then it lies on a hyperplane of interior reflection, and λ decreases as B moves perpendicularly away from any line of symmetry of Ω. The argument at the boundary is much as for case (a). The perpendicular line from the point of contact is a hyperplane of interior reflection except when the contact is at the midpoint of an edge of the polygon, but λ decreases when the point of contact is moved away from a midpoint on either side. The claim then results from possibility (b) of Corollary 2.2. $\square$

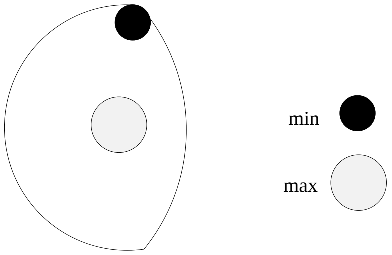
4. Extensions and some instructive examples. Although at first sight our technique seems to be restricted to regular, convex regions, we now proceed to illustrate by example how it can be extended. In this section we describe only problems with obstacles. In most cases, however, the same examples illustrate the situation for wells, with the usual reversal of "min" and "max."

We shall omit details of the proofs when they consist only of recalling Corollary 2.2 and elementary exercises in finding possible hyperplanes of interior reflection.

Example 1. Let Ω be a finite region bounded by two spheres in R^N, and suppose that a spherical obstacle B has radius ρ small enough so that it fits into Ω. For definiteness, suppose that the larger sphere has its center at the origin and radius R, and that the smaller one has its center on the x_1-axis at coordinate $x_1 = a \geq 0$, and radius $r \leq R$.

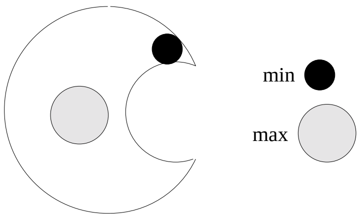
There are five possibilities as follow.

(a) Ω is gibbous (simply connected and convex: Ω is the intersection of two balls, and $R - r < a < R + r$). Then at the minimizing positions of B, it is as near as possible to a vertex. (There are two such positions if $N = 2$, and otherwise they form a sphere of dimension $N - 2$.) At the maximizing position $\mathbf{x}$ the center of B is located on the x_1-axis with x_1 in the interval $[\frac{R-r+a}{2}, \frac{R^2-r^2+a^2}{2a}] \cap [a - r + \rho, R - \rho]$. (Note: $\frac{R^2-r^2+a^2}{2a}$ is the x-coordinate of the intersection of the two circles.)



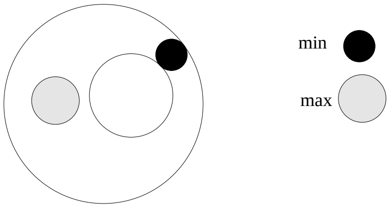
Example 1(a)

(b) Ω is crescent (simply connected and nonconvex: Ω is the intersection of the larger ball and the exterior of the smaller ball, and $R - r < a < R + r$). Then at the minimizing positions of B, it is as near as possible to a vertex. (There are two such positions if $N = 2$, and otherwise they form a sphere of dimension $N - 2$.) At the maximizing position $\mathbf{x}$ the center of B is located on the x_1-axis with x_1 in the interval $[\frac{-R-r+a}{2}, 0] \cap [-R + \rho, a - r - \rho]$.



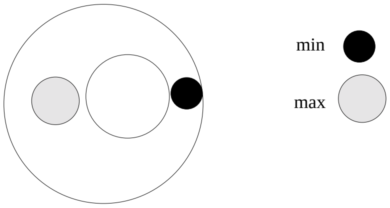
Example 1(b)

(c) Ω is annular but not concentrically so, and $\rho > \frac{R-(r+a)}{2}$. This is essentially like the case of the crescent: At the minimizing positions of B, it is as near as possible to the point $x_1 = R$ on the x_1-axis. (There are two such positions if $N = 2$, and otherwise they form a sphere of dimension $N - 2$.) The maximizing position $\mathbf{x}$ of the center of B is located on x_1-axis with x_1 in the interval $[\frac{-R-r+a}{2}, 0] \cap [-R+\rho, a-r-\rho]$, as in part (b).



Example 1(c)

(d) Ω is annular but not concentrically so, and $\rho < \frac{R-(r+a)}{2}$. The maximizing position $\mathbf{x}$ of B is as in parts (b) and (c). The minimizing positions satisfy $a+r+\rho < x_1 < \frac{a+R+r}{2}$.



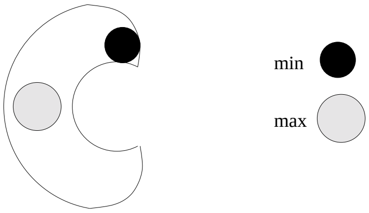
Example 1(d)

(e) Ω is a concentric annulus (spherical shell). Necessarily, $\rho < \frac{R-r}{2}$. At maximum, the position of the center $\mathbf{x}$ of B satisfies $r + \rho \leq |\mathbf{x}| \leq \frac{R+r}{2}$. At minimum, either $\mathbf{x}$ is in the same annulus or else $|\mathbf{x}| = R - \rho$ (contact with the outer boundary).

These possibilities rely on Corollary 2.2; among the hyperplanes of interior reflection to consider are horizontal hyperplanes and the bisecting planes of the spheres. Our technique does not allow us to eliminate one of the two possibilities for the minimum in 1(e). We suspect the true alternative to be that the minimizing obstacle is in contact with the outer boundary, and we will establish this in Example 12 when the inner radius is sufficiently small and the obstacle is soft.

Example 2. Horseshoe-shaped domains: Let Ω_0 be the concentric annular domain of Example 1(e) in 2 dimensions, and let S_α be the sector $|x_2| < \alpha x_1$ for some $\alpha > 0$. For some fixed opening angle $\beta \leq \pi/2$, let $\sigma_\pm$ denote two open circular sectors of radius $R-r$ and centered at the inner edge of Ω_0, i.e., at $(\frac{r}{(1+\alpha^2)^{1/2}}, \pm\frac{\alpha r}{(1+\alpha^2)^{1/2}})$, such that one edge lies on ∂S_α. $\Omega := \{\Omega_0 \cap S_\alpha^c\} \cup \sigma_- \cup \sigma_+$. The hard or soft obstacle is a disk of radius $< R-r$.

The maximizing position is on the x_1-axis and is otherwise as in Example 1(e), while the unique two minimizing positions have the obstacle wedged in the corners of the sectors $\sigma_\pm$.



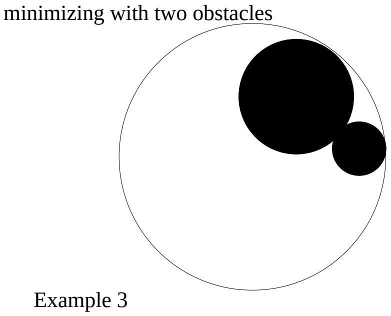
Example 2

The proof is obtained by noting that rays from the origin other than the negative x_1-axis are lines of interior reflection whenever they intersect Ω_0, as are rays beginning at $(\frac{r}{(1+\alpha^2)^{1/2}}, \pm\frac{\alpha r}{(1+\alpha^2)^{1/2}})$ and passing through $\sigma_\pm$.

We note that appending $\sigma_\pm$ allowed us to be precise about the minimizing position; without them, the minimizing position would be in contact with ∂S_α, but our method would not give the exact position.

Example 3. (a) A ball Ω with two interior hard spherical obstacles B_1 and B_2, which can be placed independently, and may have different radii, the sum of which is less than the radius of Ω. In this case the minimizing configuration has both obstacles touching the boundary and each other. In the maximizing configuration the centers of B_1 and B_2 lie on a diameter of Ω.

(b) The same example, except that one of the obstacles is allowed to be soft. The maximizing and minimizing configurations are as for part (a).

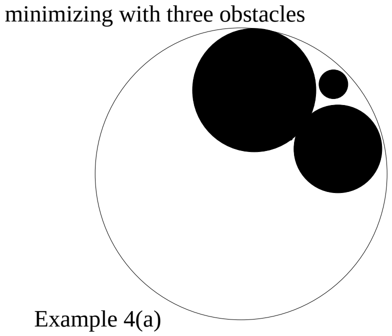


We observe that the minimizing configuration disconnects Ω.

Denote by $\mathbf{x_1}$ and $\mathbf{x_2}$ the centers of the balls B_1 and B_2, respectively. We begin by considering the relative positions of the balls as fixed a priori and treating them as a single obstacle. Unless they lie on a diameter of Ω, the line passing through both their centers is a line of interior reflection. We thus conclude that at maximum they lie on a diameter of Ω, while at minimum at least one of them touches $\partial\Omega$.

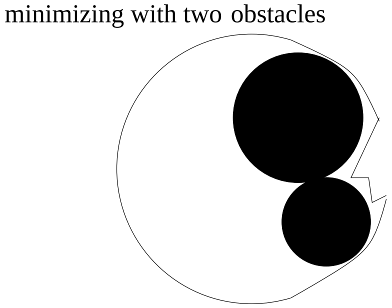
Having established that at least one obstacle touches the boundary when λ_1 is minimized, we can assume for the minimizing problem that one spherical obstacle is fixed to $\partial\Omega$ in some standard orientation, rotating the entire problem as necessary, while letting the position of the second obstacle vary. This, however, is the situation of Example 1(b). Hence we know that at minimum the two obstacles touch each other and $\partial\Omega$.

Example 4. (a) The same as Example 3, except that we also insert a third, hard obstacle of positive capacity and any shape small enough that it can be translated and rotated so as to fit inside $\Omega_s :=$ the smaller of the two domains into which Ω is disconnected at the minimizing configuration of Example 3. (Ω_1 is of course defined only up to rotations of the two larger obstacles about the center of Ω.) Then the minimizing configuration is as in Example 3, with the third obstacle anywhere within Ω_1. This can be a unique minimizer or highly nonunique, depending on the size and shape of the third obstacle.



Example 4(a)

(b) As a variant of (a), we replace the large sphere Ω with $\Omega' := \Omega \setminus \Upsilon$, where $\Upsilon \subset \Omega_1$ for some fixed possible Ω_1 as in (a), and is otherwise an arbitrary closed set of positive capacity. We consider Ω' as the exterior region, and insert two spherical obstacles. Then the minimizing configuration is as in Example 3, oriented so that Υ lies within the Ω_1 created by the disconnection of Ω. Again, this can be unique or nonunique.



Example 4(b)

Here we recall the principle of domain monotonicity, which shows that the minimal fundamental eigenvalue for Example 3 is the same as the that of the larger of the two domains into which Ω is disconnected. In case (a), if the third obstacle were inserted into this domain, the eigenvalue would strictly increase (cf. [Co95], [McG96], [McG98]). On the other hand, the eigenvalue is unaffected in comparison to Example 3 if the third obstacle is inserted into Ω_1. Case (b) follows by virtually the same argument, recalling Theorem 2.5.

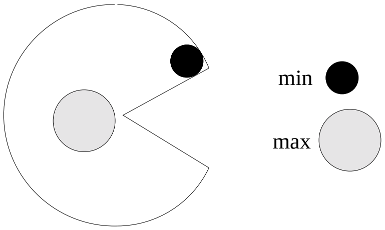
Example 5. Balls with sectors or lines (resp., hyperplane) segments removed.

(a) Consider a domain Ω formed by removing from a ball of radius R centered at the origin, some subset Υ of a closed sector S symmetric about the x_1-axis, within the half-plane $x_1 \geq a \geq 0$; otherwise, Υ is assumed closed and of positive capacity. We call the angle between the edge of the sector and the positive x_1-axis β. The obstacle is a hard or soft ball of radius $\rho < \frac{a+R}{2}$.

The maximizing position is then within the triangle (2 dimensions) or cone (3 or more dimensions) bounded by $x_1 = \frac{a-R}{2}$, $x_1 = 0$, and the cone with vertex at the origin and making an angle of $arccot(\cot\ \beta\ +\ \frac{\mathrm{a}}{\mathrm{R-a}} \csc\ \beta\ )$ with the negative x_1-axis. (Further restrictions on the maximizing position could be precisely formulated if ρ is not sufficiently small.)

The minimizing position(s) are in contact with S (possibly penetrating into its interior).

(b) As a special case of (a), suppose that Υ is a symmetric sector with vertex at $x_1 = a$, including the limiting case of a line segment. Then we can more precisely say that the maximizing position lies on the negative x_1-axis with $\frac{a-R}{2} < x_1 < 0$, and the minimizing positions are in contact with both the spherical part of $\partial\Omega$ and with ∂S. The minimizing position is not unique.

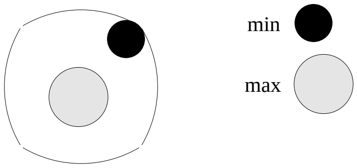


Example 5

Here we identify as hyperplanes of interior reflection all planes perpendicular to the x_1-axis with intersection at $-R < x_1 < \frac{a-R}{2}$, all planes perpendicular to the x_1-axis with intersection at $0 \leq x_1 < R$, and all planes through the origin which do not intersect S. It is a trigonometric exercise with the latter which leads to the angle identified.

As for case (b), there are additional hyperplanes of interior reflection consisting of all hyperplanes through the origin except those containing the x_1-axis.

Example 6. Let Ω be an equilateral n-sided polygon, modified by the replacement of its edges with outward circular arcs of equal angular measure $\leq \frac{2\pi}{n}$, containing a hard or soft circular obstacle B. Then the maximizing position $\mathbf{x}$ of the center of B is at the center of the polygon, and the minimizing positions of B put it as near as possible to a vertex.



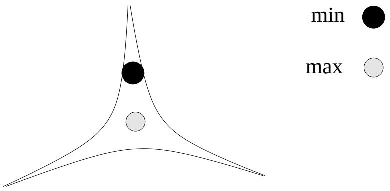
Example 6

(Straightforward exercise.)

Example 7. Let Ω be a rectangle and suppose that the convex obstacle B has two axes of symmetry parallel to the sides of the rectangle. We consider translations of the obstacle B. The maximizing position is at the center and the minimizing position is as near as possible to a vertex.

The proof is similar to the proof of Theorem 3.4.

Example 8. Let Ω be an equilateral triangle, modified by the replacement of its edges with inward circular arcs of equal angular measure $\leq \frac{\pi}{2}$, containing a hard or soft circular obstacle B. Then the maximizing position $\mathbf{x}$ of the center of B is at the center of the triangle, and the minimizing positions of B put it as near as possible to a vertex.



Example 8

(Straightforward exercise.)

Example 9. Let Ω be the concentric annulus (or spherical shell) of Example 1(e), and suppose that it contains two independently placeable hard obstacles as in Example 3(a). In the maximizing configuration, the two obstacles lie on a common hyperplane which bisects both spheres, and on opposite sides of the center. In the minimizing configuration, either the two obstacles lie on a common hyperplane which bisects both spheres, and on the same side of the center, or else they are in contact with each other (or both).

In this case, unless the obstacles are positioned as claimed, then if one of them is treated as fixed, the hyperplane passing through the origin and through the center of the second obstacle is a hyperplane of interior reflection.

Example 10. Let H be the portion of a regular helix in R^n, $n \geq 3$, at distance R from the x_1-axis for $-L \leq x_1 \leq L$. Define $\Omega := \{\mathbf{x} : \text{dist}(\mathbf{x}, H) < r\}$ for some $r < R$. (This looks like a spring with hemispherical caps on both ends.) Let the hard or soft obstacle be a ball B of radius $\rho < r$. Then the minimizing positions of B put it into contact with either of the two tips where H intersects $\partial\Omega$. At the maximizing position the center $\mathbf{x}$ of B lies on the plane perpendicular to H intersecting H at $x_1 = 0$.

Remark. Moreover, there exists $\rho_0 > 0$, depending on the specific geometry of the helix, such that for $\rho < \rho_0$ there is an upper bound on $|\mathbf{x}|$ strictly smaller than the one needed for B to fit within Ω. We do make it precise here.

The proof in this case requires a small twist in the interior reflection property on which Theorem 2.1 relies. Whereas Ω does not have the standard interior reflection property as given in the definition above Theorem 2.1, it has the following alternative property: Any hyperplane perpendicular to H other than the one intersecting it at $x_1 = 0$ divides Ω into two pieces, one of which is congruent to a subset of the other, by a half rotation instead of a reflection (equivalently, by two reflections). Applying this operation, rather than a reflection as in the proof of Theorem 2.1, we still see that the values of u and $|\nabla u|$ on one half of ∂B dominate those on the other half pointwise, allowing us to show that λ is strictly monotonic with respect to displacements tangential to H by (1.2) or, respectively, (1.3).

Our next example involves a soft obstacle and shows that our analysis is stable with respect to some perturbations of Ω which destroy both convexity and symmetry.

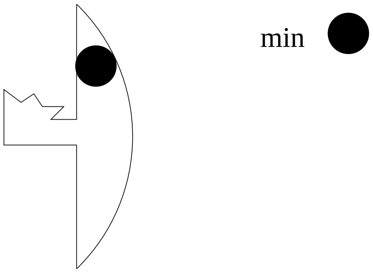
Example 11. A half ellipse with a small handle: Let $\Omega_0 := \{(x, y) \in R^2; x^2 + (y/l)^2 < 1, x > 0\}$, where l is a fixed constant, $l > 1$. Now consider the domain $\Omega_\epsilon := \mathit{interior}\,(\Omega_0 \cup H_\epsilon)$, where H_ϵ is a "handle" with the following properties:

(i) H_ϵ is a closed subset of $\{x \leq 0\}$,

(ii) $H_\epsilon \subset \{|y| \leq \epsilon\}$,

(iii) $0 < vol(H) \leq C\epsilon$ for some fixed positive constant C.

Fix the radius $r < 1$ of a spherical soft obstacle B, and denote $L_{\Omega_\epsilon} := -\Delta + \alpha\chi_B$ on Ω_ϵ. We claim that for all $\epsilon > 0$ and sufficiently small, the minimizing obstacle is situated within Ω_0 at the greatest possible distance from the x-axis. (There are two possibilities, one with $y > 0$ and one with $y < 0$.)



Example 11

Proof. We assume that ϵ is sufficiently small that the obstacle does not fit into the handle. All horizontal lines with $\{|y| > \epsilon\}$ are lines of interior symmetry for Ω_ϵ, so at the minimizing positions for the obstacle, either it is in contact with the boundary or else its center lies within $\{|y| \leq \epsilon\}$. The argument used to prove Theorem 3.2 is easily adapted for Ω_ϵ, and shows that the only possible minimizing positions are either as stated in the theorem, i.e., with the maximal value of $|y|$, or else confined to a strip of the form $\{|y| \leq c_1\epsilon\}$, where c_1 is a constant depending only on l.

We next eliminate the possible positions in the epsilonic strip by a continuity argument, for we know by the modification of Theorem 3.2 that for Ω_0, an obstacle in the strip gives rise to a fundamental eigenvalue strictly larger than when the obstacle is at the maximal value of $|y|$ (i.e., the unique two minimizing positions for Ω_0).

Consider the operator $L_\Omega := -\Delta + \alpha\chi_B$ on some domain Ω with Dirichlet boundary conditions on $\partial\Omega$ for a ball $B \subset \Omega$ of radius r, and denote, respectively, by $\lambda_\Omega(B)$ and $u_{\Omega,B}(B)(\mathbf{x})$ the first eigenvalue and L^2-normalized first eigenfunction.

Case 1. Candidate positions where B does not lie wholly inside Ω_0. In this case, according to Lemma 1.2, the obstacle can be shifted into the interior of Ω_0, raising λ by a small error. By a variant of the argument for Theorem 2.5 and uniform control on $\|u\|_\infty$ (e.g., [GiTr83, Theorem 8.15]), this error tends to 0 as $\epsilon \to 0$ uniformly in these possible candidate positions. Hence it suffices to consider only candidate positions inside Ω_0.

Case 2. Candidate positions inside Ω_0 and inside the epsilonic strip $\{|y| \leq c_1\epsilon\}$. Here estimates as in [HiMa91], [GeZh94], [McG96], [McG98] show that $\lambda_{\Omega_\epsilon}(B) \to \lambda_{\Omega_0}(B)$ as $\epsilon \to 0$.

Example 12. Example 1(e) revisited: Ω is a concentric annulus (spherical shell) of outer radius R and inner radius r. In it we place a soft spherical obstacle centered at $\mathbf{w}$ and of radius ρ; as in 1(e), necessarily, $\rho < \frac{R-r}{2}$. If r is sufficiently small, then at the minimizing position the support $|\mathbf{w}| = R - \rho$, i.e., the obstacle touches the outer boundary of the annulus.

The strategy of the proof is rather general and can be applied to some other cases as indicated below. For this reason we discuss it in some detail. For simplicity, we discuss the case of two dimensions.

Fix the radius ρ of the obstacle and the coupling α. We already know that $\mathbf{w} \in \mathcal{C}_1 \cup \mathcal{C}_2$, where

$$\mathcal{C}_1 := \{r + \rho \leq |\mathbf{x}| \leq (R + r)/2\}, \quad \mathcal{C}_2 = \{|\mathbf{x}| = R - \rho\}.$$

Note that $R/2 < R - \rho$, and

$$\mathcal{C}_1 \subset \mathcal{C}_1^* \equiv \{\rho \leq |\mathbf{x}| \leq R/2\}.$$

We denote by Λ and Λ^r the minima of the fundamental eigenvalues given a soft obstacle of radius ρ in domains $\Omega_0 = \{|x| < R\}$ and Ω_r, respectively.

Now we claim the following.

Claim A. We have $|\Lambda - \Lambda^r| \leq C/|\log r|$, where C is a constant independent of r.

Indeed, denoting by $B_*^r = B_{*,\rho}^r$ the optimal ball for the problem on Ω_r, we have $\Lambda^\epsilon = \lambda_{\Omega_r}(B_*^r) \geq \lambda_{\Omega_0}(B_*^r) \geq \Lambda$ because of domain monotonicity. Here $\lambda_\Omega(B)$ is the first eigenvalue of $-\Delta + \alpha\chi_B$ on Ω. On the other hand, denoting by $B_* = B_{*,\rho}$ the optimal ball for the problem on Ω_0, we know by [Sw63, Theorem 2], for example, that

$$\lambda_{\Omega_r}(B_*) \leq \lambda_{\Omega_r}(B_*) + \frac{C}{|\log r|}.$$

This yields

$$\Lambda^r \leq \Lambda + \frac{C}{|\log r|},$$

which implies Claim A.

We know that for $B_*^r(\mathbf{w}(r))$,

$$\rho + r \leq |\mathbf{w}(r)| \leq (R + r)/2 \quad \text{ or } \quad |\mathbf{w}(r)| = R - \rho.$$

If the conclusion did not hold, then there would exist a subsequence $\{r_j\}, r_j \to 0$ and $B^{r_j}(\mathbf{w}(r_j))$, such that

$$\rho + r_j \leq |\mathbf{w}(r_j)| \leq (R + r_j)/2.$$

By passing if necessary to a further subsequence, we may assume that $\mathbf{w}(r_j)$ converges to $\mathbf{w}_0$ with $|\mathbf{w}_0| \in [\rho, R/2]$. We use the notation

$$B_0 = B_\rho(w_0), \quad B_*^j = B_*^{r_j}(w(r_j)),$$

for simplicity. Next we claim the following.

Claim B. As $j \to \infty$, $\lambda_{\Omega_{r_j}}(B_*^j) - \lambda_{\Omega_{r_j}}(B_0) \to 0$.

Again using [Sw63], we obtain

$$|\lambda_{\Omega_{r_j}}(B_0) - \lambda_{\Omega_0}(B_0)| \to 0.$$

Granting (B), we get

$$|\lambda_{\Omega_{r_j}}(B_*^j) - \lambda_{\Omega_0}(B_0)| \leq |\lambda_{\Omega_{r_j}}(B_*^j) - \lambda_{\Omega_{r_j}}(B_0)| + |\lambda_{\Omega_{r_j}}(B_0) - \lambda_{\Omega_0}(B_0)| \to 0.$$

Combining this with (A), we conclude that

$$\lambda_{\Omega_0}(B_0) = \Lambda,$$

which contradicts the definition of B_0 and the stated fact about Λ. Thus, we conclude the desired statement.

It remains only to prove Claim B.

We shall show that

$$|\lambda_{\Omega_{r_j}}(B_*^j) - \lambda_{\Omega_{r_j}}(B_0)| \leq \alpha M |B_*^j \ominus B_0|^{1/2},$$

where $A \ominus B = (A \setminus B) \cup (B \setminus A)$. As a consequence we obtain $\lambda_{\Omega_{r_j}}(B_*^j) - \lambda_{\Omega_{r_j}}(B_0) \to 0$.

First, by definition, $\lambda_{\Omega_{r_j}}(B_*^j) \leq \lambda_{\Omega_{r_j}}(B_0)$. Denote by ϕ_{r_j} the normalized eigenfunction associated with $\lambda_{\Omega_{r_j}}(B_*^j)$. Then

$$\begin{aligned}
\lambda_{\Omega_{r_j}}(B_0) &\leq \int_{\Omega_{r_j}} |\nabla \phi_{r_j}|^2 + \alpha \chi_{B_0} \phi_{r_j}^2 \\
&= \int |\nabla \phi_{r_j}|^2 + \alpha \chi_{(B_*^j)} \phi_{r_j}^2 \, dx + \alpha \left(\int_{B_0} \phi_{r_j}^2 \, dx - \int_{B_*^j} \phi_{r_j}^2 \, dx \right) \\
&\leq \lambda_{\Omega_{r_j}}(B_*^j) + \alpha \int_{B_0 \ominus (B_*^j)} \phi_{r_j}^2 \, dx \\
&\leq \lambda_{\Omega_{r_j}}(B_*^j) + \alpha \|\phi_{r_j}\|_{L^4(\Omega_{r_j})}^2 |B_0 \ominus (B_*^j)|^{1/2}.
\end{aligned}$$

It is easy to see that $\lambda_{\Omega_{r_j}}(B_*^j)$ is bounded, and hence that $\int_{\Omega_{r_j}} |\nabla \phi_{r_j}|^2 \leq M$. By the Sobolev embedding theorem, we obtain $\|\phi_{r_j}\|_{L^4(\Omega_{r_j})}^2 \leq M$, yielding the desired estimate.

The strategy of the perturbation argument used for Example 12 can be used in many other situations. To summarize, let Ω_ϵ be a (singular) perturbation of Ω and assume, for simplicity, $\Omega_\epsilon \subset \Omega$. For a fixed $D \subset \Omega$, assume that $\lambda^\epsilon(D) \to \lambda(D)$ as $\epsilon \to 0$. Here $\lambda^\epsilon(D)$ and $\lambda(D)$ are the fundamental eigenvalues of $-\Delta + \alpha \chi_D$ on Ω_ϵ and Ω, respectively. We denote by $\Lambda = \lambda(B(x_0))$ and $\Lambda_\epsilon = \lambda(B(x_\epsilon))$ the optimal eigenvalues on Ω_ϵ and Ω, respectively. Then, if $x_0 \in \mathcal{C}_0$ and $x_\epsilon \in \mathcal{C}_0 \cup \mathcal{C}_\epsilon$ with $\mathcal{C}_\epsilon \subset \mathcal{C}$, and if $\inf_{y \in \mathcal{C}} \lambda(B(y)) > \Lambda$, then $x_\epsilon \in \mathcal{C}_0$ for sufficiently small ϵ.

We close by stating another perturbative result related to Examples 5 and 12.

Example 13. In two dimensions, let $\Omega_\epsilon = \{\epsilon < |x| < b\} \setminus \{(r, \theta); \epsilon < r < b, |\theta| \leq \beta\}$ with $\beta < \pi/2$, and suppose that it contains a soft spherical obstacle. Then for sufficiently small ϵ, at the minimizing position the obstacle is as close as the corner points $(r, \theta) = (b, \pm\beta)$.

Acknowledgments. This collaboration was made possible by two visiting invitations, that of Pawel Kröger to the Georgia Institute of Technology in Atlanta in April, 1995, and that of Evans Harrell and Kazuhiro Kurata to the Erwin Schrödinger Institut in Vienna in May, 1998. We are grateful to those institutions, as well as the French C.N.R.S. and the University of Toulouse 1 and 3, for their hospitality. Moreover, we wish to thank Ivor McGillivray for useful references.

REFERENCES

[Al60] A. D. Alexandrov, *Certain estimates for the Dirichlet problem*, Soviet Math. Dokl., 1 (1960), pp. 1151–1154.

[As99] M. S. Ashbaugh, *private communication.*

[BeNi91] H. Berestycki and L. Nirenberg, *On the method of moving planes and the sliding method*, Bol. Soc. Brasil Mat. (N.S.), 22 (1991), pp. 1–37.

[CGIKO99] S. Chanillo, D. Grieser, M. Imai, K. Kurata, and I. Ohnishi, *Symmetry breaking and other phenomena in the optimization of eigenvalues for composite membranes*, Comm. Math. Phys., 214 (2000), pp. 315–337.

[Co94] A. Colesanti, *A variational problem for harmonic functions in ring-shaped domains with partially free boundary*, SIAM J. Math. Anal., 25 (1994), pp. 1122–1127.

[Co95] G. Courtois, *Spectrum of manifolds with holes*, J. Funct. Anal., 134 (1995), pp. 194–221.

[Da95] E. B. Davies, *Spectral Theory and Differential Operators*, Cambridge Stud. Adv. Math. 42, Cambridge University Press, Cambridge, UK, 1995.

[Da95a] E. B. Davies, *private communication*, 1995.

[Fl93] M. Flucher, *An asymptotic formula for the minimal capacity among sets of equal area*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 71–86.

[Fl95] M. Flucher, *Approximation of Dirichlet eigenvalues on domains with small holes*, J. Math. Anal. Appl., 193 (1995), pp. 169–199.

[GaSc53] P. R. Garabedian and M. Schiffer, *Convexity of domain functionals*, J. Analyse Math., 2 (1953), pp. 281–368.

[GeZh94] F. Gesztesy and Z. Zhao, *Domain perturbations, Brownian motion, capacities, and ground states of Dirichlet Schrödinger operators*, Math. Z., 215 (1994), pp. 143–150.

[GiTr83] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Grundlehren Math. Wiss. 224, Springer-Verlag, Berlin, New York, 1983.

[Ha08] J. Hadamard, *Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrées*, Méms. prés. par div. savants a l'Acad. des Sci., 33 (1908), pp. 1–128.

[HiMa91] P. D. Hislop and A. Martinez, *Scattering resonances of a Helmholtz resonator*, Indiana Univ. Math. J., 40 (1991), pp. 767–788.

[JeKe85] D. Jerison and C. Kenig, *Unique continuation and absence of positive eigenvalues for Schrödinger operators*, Ann. Math., 121 (1985), pp. 463–494.

[McG96] I. McGillivray, *Capacitary estimates for Dirichlet eigenvalues*, J. Funct. Anal., 139 (1996), pp. 244–259.

[McG98] I. McGillivray, *Capacitary asymptotic expansion of the groundstate to second order*, Comm. Partial Differential Equations, 23 (1998), pp. 2219–2252.

[PrWe84] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Springer-Verlag, New York, 1984.

[ReSi78] M. Reed and B. Simon, *Methods of Modern Mathematical Physics,* IV. *Analysis of Operators*, Academic Press, New York, 1978.

[Se71] J. Serrin, *A symmetry problem in potential theory*, Arch. Rational Mech. Anal., 43 (1971), pp. 304–318.

[Sw63] C. A. Swanson, *Asymptotic variational formulae for eigenvalues*, Canad. Math. Bull., 6 (1963), pp. 15–25.

SIAM J. MATH. ANAL.
Vol. 33, No. 2, pp. 261–285

AN EXACTLY SOLVABLE MODEL FOR THE INTERACTION OF LINEAR WAVES WITH KORTEWEG–DE VRIES SOLITONS[*]

P. D. MILLER[†] AND S. R. CLARKE[‡]

Abstract. Under certain mode-matching conditions, small-amplitude waves can be trapped by coupling to solitons of nonlinear fields. We present a model for this phenomenon, consisting of a linear equation coupled to the Korteweg–de Vries (KdV) equation. The model has one parameter, a coupling constant κ. For one value of the coupling constant, $\kappa = 1$, the linear equation becomes the linearized KdV equation, for which the linear waves can indeed be trapped by solitons and, moreover, for which the initial value problem for the linear waves has been solved exactly by Sachs [S83] in terms of quadratic forms in the Jost eigenfunctions of the associated Schrödinger operator. We consider in detail a different case of weaker coupling, $\kappa = 1/2$. We show that in this case linear waves may again be trapped by solitons, and like the stronger coupling case $\kappa = 1$, the initial value problem for the linear waves can also be solved exactly, this time in terms of linear forms in the Jost eigenfunctions. We present a family of exact solutions, and we develop the completeness relation for this family of exact solutions, finally giving the solution formula for the initial value problem. For $\kappa = 1/2$, the scattering theory of linear waves trapped by solitons is developed. We show that there exists an explicit increasing sequence of bifurcation values of the coupling constant, $\kappa = 1/2, 1, 5/3, \ldots$, for which some linear waves may become trapped by solitons. By studying a third-order eigenvalue equation, we show that for $\kappa < 1/2$ all linear waves are scattered by solitons, and that for $1/2 < \kappa < 1$, as well as for $\kappa > 1$, some linear waves are *amplified* by solitons.

Key words. solitons, Korteweg–de Vries equation, coupled systems, completeness relations, wave trapping

AMS subject classifications. 37K40, 35Q53, 42A65

PII. S0036141099365431

1. Introduction.

This paper is concerned with solving the coupled system of equations

$$\partial_t A + \partial_x \left[\frac{1}{2}A^2 + \partial_x^2 A\right] = 0, \tag{1}$$

$$\partial_t B + \partial_x \left[\kappa A B + \partial_x^2 B\right] = 0\,, \tag{2}$$

where κ is a real parameter. Of course, the nonlinear equation for $A(x,t)$ is simply the Korteweg–de Vries (KdV) equation, and it can be solved independently by the inverse-scattering transform [GGKM67]. The coupled system (1) and (2) is a partially linearized version of the system proposed by Hirota and Satsuma [HS81] as a model for the dynamics of coupled long waves.

The coupled system (1) and (2) can be solved exactly when $\kappa = 1$ and when $\kappa = 1/2$. The case of $\kappa = 1$ is well known, for then the equation (2) is just the KdV equation itself linearized about the solution $A(x,t)$. An elementary exact solution of the linear equation (2) in this case is given by $B(x,t) = \partial_x A(x,t)$. Further solutions can be expressed in terms of derivatives of the squared eigenfunctions of the related Schrödinger operator with potential A [S83].

[*] Received by the editors December 7, 1999; accepted for publication November 27, 2000; published electronically June 8, 2001.

http://www.siam.org/journals/sima/33-2/36543.html

[†] Department of Mathematics, University of Michigan, East Hall, 525 East University Avenue, Ann Arbor, MI 48109 (millerpd@umich.edu).

[‡] Department of Mathematics and Statistics, Monash University, P.O. Box 28M, Victoria 3800, Australia (src@mail.maths.monash.edu.au).

The case of $\kappa = 1/2$ is essentially different. In this case, the linear equation (2) is no longer the linearization of KdV about *any* solution. An elementary exact solution of the linear equation in this case is given simply by $B(x, t) = A(x, t)$. The main goal of this paper is to construct the *general* solution of the initial value problem for this linear equation when $A(x, t)$ is a multisoliton solution of KdV.

One way to make clear the difference between the cases $\kappa = 1$ and $\kappa = 1/2$ is to consider $A(x, t)$ to be the simple soliton solution of KdV (1):

$$A(x,t) = 12\eta^2 \mathrm{sech}^2(\eta(x - 4\eta^2 t - \alpha)) = -V(\chi)\,, \tag{3}$$

where $\chi = x - ct - \alpha$ and the velocity is $c = 4\eta^2$. If we look for solutions of the linear equation (2) that are traveling waves with speed c, we find the equation

$$[-\kappa V(\chi)B(\chi) + B''(\chi)]' = cB'(\chi)\,. \tag{4}$$

Integrating once, using vanishing boundary conditions at $\chi = \pm\infty$, yields a Schrödinger eigenvalue problem for B:

$$-B''(\chi) + \kappa V(\chi)B(\chi) = EB(\chi)\,, \tag{5}$$

where $E = -c$. For $\kappa = 1/2$, it follows from the fact that $B(x, t) = A(x, t)$ is a solution of (2) that the function $B(\chi) = V(\chi)$ is an eigenfunction of the Schrödinger operator with eigenvalue $E = -c = -4\eta^2$. Since it has no zeros, it is the ground state eigenfunction. We will see below that there is also one excited state for $\kappa = 1/2$, although it is not relevant here since it corresponds to a different velocity. On the other hand, for $\kappa = 1$, $B(x, t) = \partial_x A(x, t)$ is a solution of (2), which implies that the function $B(\chi) = \partial_x V(\chi)$ is an eigenfunction of the Schrödinger operator with the same eigenvalue $E = -c = -4\eta^2$. In this case, the eigenfunction has a single zero and therefore is the first excited state. It follows that there are at least two eigenvalues for $\kappa = 1$. In fact, there are exactly three states in this case. A final observation is that from the construction of the one-soliton solution of KdV (see (8), (9), and (10) below) it follows that for $\kappa = 1/6$, the function $B(\chi) = V(\chi)$ is an eigenfunction of the Schrödinger operator with eigenvalue $E = -c/4 = -\eta^2$. It is the ground state and the only eigenfunction. These relationships are summarized in Figure 1.1. We will have more to say about this picture when we discuss the trapping of linear waves by solitons for general values of κ in section 6.

The rest of this paper is primarily concerned with developing the general solution of the initial value problem for (2) with $\kappa = 1/2$ when $A(x, t)$ is an N-soliton solution of KdV (1). In section 2 we show how for $\kappa = 1/2$ a large family of exact solutions of (2) can be obtained from the simultaneous solutions of the Lax pair for KdV. When the solution of KdV contains only solitons and no radiation, the construction of Lax eigenfunctions is completely algorithmic and algebraic, and consequently the corresponding family of solutions of (2) for $\kappa = 1/2$ can be obtained with great practicality. In section 3 we then establish that in the N-soliton case there are enough of these exact solutions of (2) for $\kappa = 1/2$ to expand for fixed t any absolutely continuous $L^1(\mathbb{R})$ function of x. This fact then leads to a general solution formula for (2) simply by expanding the initial data. We present and discuss this formula in section 4. There will turn out to be N linearly independent solutions that are asymptotically confined to the union of soliton trajectories and can therefore be considered to be bound states. In section 5 we compute the scattering matrix that relates the asymptotic behavior of bound states for $t \to -\infty$ to the corresponding behavior for $t \to +\infty$. In section 6 we

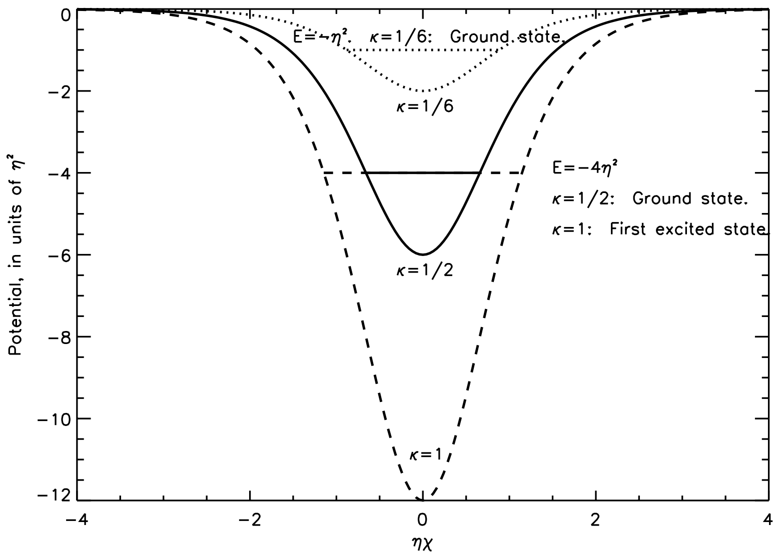


FIG. 1.1. *Energy levels of the* $12\kappa\eta^2\mathrm{sech}^2(\eta\chi)$ *potential for three different values of the coupling constant* κ.

consider general values of the coupling constant κ and describe the behavior of some solutions of (2) when $A(x,t)$ is a one-soliton solution of KdV (1). These calculations indicate the exceptional nature of the two values $\kappa = 1/2$ and $\kappa = 1$. In the appendix, we describe several physical applications of the coupled system (1) and (2) to topics in molecular dynamics, mechanics, soliton theory, and the fluid dynamics of internal waves.

2. Exact solution formulas for $\kappa = 1/2$. As is well known [GGKM67], the KdV equation (1) is the compatibility condition for a pair of linear equations involving a complex parameter λ for an auxiliary function $f(x,t,\lambda)$. This pair of linear equations is

$$(6) \qquad \partial_x^2 f = -\frac{\lambda^2}{4} f - \frac{1}{6} A f \qquad \text{and} \qquad \partial_t f = \frac{1}{6}\partial_x A \cdot f + \left(\lambda^2 - \frac{1}{3}A\right)\partial_x f$$

and is called a Lax pair. A simultaneous solution $f(x,t,\lambda)$ of these linear equations exists if and only if the function $A(x,t)$ satisfies KdV (1). Suppose that this is the case. Then, it is a direct matter to verify that for fixed but arbitrary $\lambda \in \mathbb{C}$, the two functions defined by

$$(7) \qquad B(x,t) := \partial_x \left[f(x,t,\lambda) \exp\left(\pm \frac{i}{2}(\lambda x + \lambda^3 t) \right) \right]$$

are solutions of the linear equation (2) when $\kappa = 1/2$. Note here an important point of departure from the other solvable case, namely, $\kappa = 1$, where (2) is the linearized KdV equation. In the latter case, particular solutions are expressed in terms of the x-derivative of the *square* of the Lax eigenfunction $f(x,t,\lambda)$ [GGKM74, S83]. By contrast, the formula (7) for solutions of (2) for $\kappa = 1/2$ is linear in $f(x,t,\lambda)$. This fact leads to some important simplifications.

The formula (7) is only really practical to use if one can explicitly compute the function $f(x,t,\lambda)$. This will be the case if the solution $A(x,t)$ of KdV (1) is a pure N-soliton solution. For each fixed t, $A(x,t)$ is then a reflectionless potential of the Schrödinger equation in the Lax pair (6). The multisoliton solutions of KdV and the associated solutions of the Lax pair are constructed as follows [KM56]. Let $f_+(x,t,\lambda)$ be given by

$$f_+(x,t,\lambda) := \left(1+\sum_{n=0}^{N-1}\lambda^{n-N}f_n(x,t)\right)\exp\left(-\frac{i}{2}(\lambda x+\lambda^3 t)\right)\,, \tag{8}$$

where the $f_n(x,t)$ are unknown coefficients. Choose N positive numbers $\eta_1 > \eta_2 > \cdots > \eta_N$, and N arbitrary real numbers $\alpha_1,\ldots,\alpha_N$, and insist that $f_+(x,t,\lambda)$ satisfy the relations

$$f_+(x,t,2i\eta_n) = (-1)^{n+1}\exp(2\eta_n\alpha_n)f_+(x,t,-2i\eta_n) \tag{9}$$

for all $n=1,\ldots,N$. It is easy to see that these relations imply a square linear algebraic system for the coefficients $f_n(x,t)$. The determinant of the system is always nonzero, and so the coefficients $f_n(x,t)$ are determined uniquely from the soliton eigenvalues $\{\lambda_n = 2i\eta_n\}$ and the norming constants $\{\alpha_n\}$ in terms of exponential functions. From this construction, it can be shown that if one chooses

$$A(x,t) := 6i\partial_x f_{N-1}(x,t)\,, \tag{10}$$

then $f_+(x,t,\lambda)$ and $f_-(x,t,\lambda) := f_+(x,t,-\lambda)$ are two simultaneous solutions of the Lax pair (6), and the function $A(x,t)$ defined by (10) satisfies KdV (1). The two functions $f_\pm(x,t,\lambda)$ are linearly independent for all nonzero $\lambda \neq \pm 2i\eta_n$. According to the linear relations (9) that determine the coefficients, at the exceptional values of λ the two functions are proportional.

The solution $A(x,t)$ of KdV so constructed represents the interaction of N solitons. In particular, as $t \to \pm\infty$, the solution can be represented in the form

$$A(x,t) \sim \sum_{n=1}^{N} A_n^\pm(x,t)\,, \qquad \text{where} \qquad A_n^\pm(x,t) := 12\eta_n^2\operatorname{sech}^2(\eta_n(x-\alpha_n^\pm)-4\eta_n^3 t)\,, \tag{11}$$

where the asymptotic phase constants $\alpha_n^\pm$ are functions of the η_n and α_n.

Below we will need the asymptotic behavior of the functions $f_\pm(x,t,\lambda)$ as $x \to \pm\infty$ for λ and t fixed. It can be shown that the coefficient functions $f_n(x,t)$ remain bounded as $x \to \pm\infty$. Then, letting x tend to $\pm\infty$ in the linear relations (9), one finds from dominant balance arguments that

$$\lim_{x\to\pm\infty}\left(1+\sum_{n=0}^{N-1}\lambda^{n-N}f_n(x,t)\right)\Bigg|_{\lambda=\pm 2i\eta_n} = 0\,. \tag{12}$$

These relations imply that

$$\lim_{x\to\pm\infty}\left(1+\sum_{n=0}^{N-1}(\pm\lambda)^{n-N}f_n(x,t)\right) = \lambda^{-N}\prod_{n=1}^{N}(\lambda-2i\eta_n)\,. \tag{13}$$

Therefore, for all $\lambda\in\mathbb{C}$,

$$\lim_{x\to\pm\infty} f_+(x,t,\lambda)\exp\left(\frac{i}{2}(\lambda x+\lambda^3 t)\right) = \lambda^{-N}\prod_{n=1}^{N}(\lambda\mp 2i\eta_n)\,. \tag{14}$$

The large $|x|$ asymptotics for the other solution $f_-(x,t,\lambda)$ follow from the definition $f_-(x,t,\lambda) = f_+(x,t,-\lambda)$. From these asymptotics, it is easy to construct the appropriate linear combinations of $f_\pm(x,t,\lambda)$ that correspond to the Jost functions of the Schrödinger equation, normalized at $x = \pm\infty$.

In the formula (7) we have a choice of sign in the exponent. In fact, it is easy to see that if one considers the totality of solutions obtained for all complex λ, the choice of sign is redundant. In what follows, we adopt a particular choice of the sign and maintain generality by using both Lax eigenfunctions $f_+(x,t,\lambda)$ and $f_-(x,t,\lambda)$. Thus, the particular solutions of the linear equation (2) for $\kappa = 1/2$ that we will consider below will be denoted by $h_\pm(x,t,\lambda)$, given by

$$h_\pm(x,t,\lambda) := \partial_x g_\pm(x,t,\lambda)\,, \qquad \text{where} \qquad g_\pm(x,t,\lambda) := f_\pm(x,t,\lambda)\exp\left(\frac{i}{2}(\lambda x + \lambda^3 t)\right)\,. \tag{15}$$

From the Schrödinger equation (6) for $f_\pm(x,t,\lambda)$, it follows that $g_\pm(x,t,\lambda)$ satisfies the ODE

$$-i\partial_x^2 g_\pm - i\frac{A}{6}g_\pm = \lambda\partial_x g_\pm\,. \tag{16}$$

This ODE plays an important role in suggesting the completeness relation for the solutions $h_\pm(x,t,\lambda)$.

3. The completeness relation for $\boldsymbol{\kappa = 1/2}$. Having in hand a large family of exact solutions of the linear equation (2) for $\kappa = 1/2$ is certainly useful, but we may then ask whether there are enough of these solutions to construct the general solution of the initial value problem by superposition. A *completeness relation* is a formula that gives the expansion of arbitrary initial data in terms of such a collection of functions. In this section, we will establish the completeness relation for the exact solutions $h_\pm(x,t,\lambda)$ obtained in section 2.

The form of the completeness relation is suggested by a similar argument to that used by Sachs [S83] in his investigation of the completeness of squared eigenfunction solutions to the linearized KdV equation. The idea is that ideally we would like to have a differential eigenvalue problem in standard form satisfied by the functions $h_\pm(x,t,\lambda)$:

$$L(t)h_\pm(x,t,\lambda) = \lambda h_\pm(x,t,\lambda)\,, \tag{17}$$

where λ is the eigenvalue and $L(t)$ is some second-order linear differential operator in x. Then, using the two explicit solutions $h_\pm(x,t,\lambda)$ of this problem, we could solve the inhomogeneous problem

$$L(t)\psi - \lambda\psi = \phi \tag{18}$$

by variation of parameters, i.e., by writing ψ as a linear combination of $h_\pm(x,t,\lambda)$ with nonconstant coefficients, and substituting into (18). For a fixed function $\phi(x)$, this determines $\psi(x,t,\lambda)$, and we have thus constructed the resolvent of the operator $L(t)$,

$$\psi(x,t,\lambda) = (L(t) - \lambda\mathbb{I})^{-1}\phi(x)\,. \tag{19}$$

If the spectrum of $L(t)$ is contained in a bounded region of the complex plane (and also under some milder conditions), then the Dunford–Taylor integral of the resolvent on a positively oriented contour enclosing the spectrum yields the identity operator

$$-\frac{1}{2\pi i}\oint \psi(x,t,\lambda)\,d\lambda = -\frac{1}{2\pi i}\oint (L(t)-\lambda\mathbb{I})^{-1}\phi(x)\,d\lambda = \phi(x)\,. \tag{20}$$

However, we do not have a second-order eigenvalue problem for $h_\pm(x,t,\lambda)$. Instead we have the second-order equation (16) for $g_\pm(x,t,\lambda)$. However, we make the *guess* that a similar procedure will apply here. Namely, for appropriate side conditions (see below) we solve the inhomogeneous equation

$$-i\partial_x^2\psi - i\frac{A}{6}\psi - \lambda\partial_x\psi = \phi \tag{21}$$

for $\psi(x,t,\lambda)$ using variation of parameters with the two functions $g_\pm(x,t,\lambda)$ solving the homogeneous equation (16), and then we differentiate the resulting formula with respect to x. Formally speaking only, we have thus constructed the resolvent of the "operator"

$$L(t) = -i\partial_x - i\frac{A}{6}\partial_x^{-1}\,. \tag{22}$$

The obstruction to rigor here is that ∂_x^{-1} is not well defined. Nonetheless, we are guided to hypothesize that

$$-\frac{1}{2\pi i}\oint \partial_x\psi(x,t,\lambda)\,d\lambda = \phi(x) \tag{23}$$

for an appropriate contour of integration. This formula turns out to be correct, although a direct proof must be supplied. The proof we use follows Miller and Akhmediev [MA98].

3.1. Solving the inhomogeneous problem. We express the solution of the inhomogeneous problem in the form

$$\psi(x,t,\lambda) = C_+(x,t,\lambda)g_+(x,t,\lambda) + C_-(x,t,\lambda)g_-(x,t,\lambda)\,, \tag{24}$$

subject to the usual "reduction of order" condition

$$\partial_x C_+(x,t,\lambda)\cdot g_+(x,t,\lambda) + \partial_x C_-(x,t,\lambda)\cdot g_-(x,t,\lambda) = 0\,. \tag{25}$$

Substituting (24) into the equation for ψ, and using (25), one finds

$$\partial_x C_+(x,t,\lambda) = -i\frac{\phi(x)g_-(x,t,\lambda)}{W(g_+,g_-)} \qquad \text{and} \qquad \partial_x C_-(x,t,\lambda) = i\frac{\phi(x)g_+(x,t,\lambda)}{W(g_+,g_-)}\,, \tag{26}$$

where $W(g_+,g_-) := g_+\partial_x g_- - g_-\partial_x g_+$ is the Wronskian.

From the differential equation (16) satisfied by $g_\pm(x,t,\lambda)$, it follows that

$$\partial_x W(g_+,g_-) = i\lambda W(g_+,g_-)\,. \tag{27}$$

Using the large $|x|$ asymptotics of $f_\pm(x,t,\lambda)$ obtained in section 2, one then solves (27) uniquely and finds that

$$W(g_+,g_-) = i\lambda^{1-2N}\exp(i(\lambda x+\lambda^3 t))\prod_{n=1}^{N}(\lambda^2+4\eta_n^2)\,. \tag{28}$$

In solving the inhomogeneous equation (21) for ψ, we really should impose appropriate side conditions. Here, the side conditions we use are not related to boundary conditions in x as much as to analyticity conditions in λ. It is easy to check that for each $x_{0,U}$, the function $\psi_U(x,t,\lambda)$ defined by

$$\begin{aligned}(29)\qquad \psi_U(x,t,\lambda) \;&=\; -\int_{x_{0,U}}^{x} \frac{g_-(z,t,\lambda)\exp(-i(\lambda z+\lambda^3 t))\phi(z)}{\lambda^{1-2N}\prod_{n=1}^{N}(\lambda^2+4\eta_n^2)}\,dz\cdot g_+(x,t,\lambda)\\ &\quad+\int_{-\infty}^{x} \frac{g_+(z,t,\lambda)\exp(-i(\lambda z+\lambda^3 t))\phi(z)}{\lambda^{1-2N}\prod_{n=1}^{N}(\lambda^2+4\eta_n^2)}\,dz\cdot g_-(x,t,\lambda)\end{aligned}$$

is a solution analytic in λ for $\Im(\lambda)>0$ and $|\lambda|$ sufficiently large. Similarly, $\psi_L(x,t,\lambda)$ defined for each $x_{0,L}$ by

$$\begin{aligned}(30)\qquad \psi_L(x,t,\lambda) \;&=\; -\int_{x_{0,L}}^{x} \frac{g_-(z,t,\lambda)\exp(-i(\lambda z+\lambda^3 t))\phi(z)}{\lambda^{1-2N}\prod_{n=1}^{N}(\lambda^2+4\eta_n^2)}\,dz\cdot g_+(x,t,\lambda)\\ &\quad-\int_{x}^{\infty} \frac{g_+(z,t,\lambda)\exp(-i(\lambda z+\lambda^3 t))\phi(z)}{\lambda^{1-2N}\prod_{n=1}^{N}(\lambda^2+4\eta_n^2)}\,dz\cdot g_-(x,t,\lambda)\end{aligned}$$

is a solution analytic for $\Im(\lambda)<0$ and $|\lambda|$ sufficiently large. The qualification of $|\lambda|$ being sufficiently large is necessary because the expressions have poles at the soliton eigenvalues in the respective half-planes where the two functions $g_\pm(x,t,\lambda)$ become proportional. However, these are the only finite singularities, and both solutions $\psi_U(x,t,\lambda)$ and $\psi_L(x,t,\lambda)$ are meromorphic in the whole of their respective open half-planes.

The arbitrariness of the parameters $x_{0,U}$ and $x_{0,L}$ would seem to be a problem; however, it will turn out that these terms contribute nothing to the Dunford–Taylor integral that we will prove gives the required completeness relation.

3.2. Integrating the resolvent.

Here we show that the guess we made is indeed correct.

THEOREM 3.1. *Let $\phi(x)$ be an absolutely continuous function in $L^1(\mathbb{R})$. Let $x_{0,U}$ and $x_{0,L}$ be constants, and let $t\in\mathbb{R}$ be fixed. Then,*

$$(31)\qquad \phi(x) = -\frac{1}{2\pi i}\lim_{R\to\infty}\left[\int_{C_U}\partial_x\psi_U(x,t,\lambda)\,d\lambda+\int_{C_L}\partial_x\psi_L(x,t,\lambda)\,d\lambda\right],$$

where C_U is the positively oriented half-circle from R to $-R$ in the upper half-plane and C_L is the positively oriented half-circle from $-R$ to R in the lower half-plane.

Proof. First, we show that the terms depending on the arbitrary parameters $x_{0,U}$ and $x_{0,L}$ converge to zero as $R\to\infty$. This will justify calling the function ψ_U or ψ_L

a "resolvent" even though the inverse is not unique. Consider the integral

$$
\begin{aligned}
J_U &:= \int_{C_U} \partial_x \left[-\int_{x_{0,U}}^{x} \frac{g_-(z,t,\lambda)\exp(-i(\lambda z+\lambda^3 t))\phi(z)}{\lambda^{1-2N}\prod_{n=1}^{N}(\lambda^2+4\eta_n^2)}\,dz\cdot g_+(x,t,\lambda)\right] d\lambda \\
&= -\int_{C_U} \frac{\lambda^{2N}h_+(x,t,\lambda)}{\lambda\prod_{n=1}^{N}(\lambda^2+4\eta_n^2)}\int_{x_{0,U}}^{x} g_-(z,t,\lambda)\exp(-i(\lambda z+\lambda^3 t))\phi(z)\,dz\,d\lambda\,,
\end{aligned}
\tag{32}
$$

where we have used the relation (25). Recall that

$$
\begin{aligned}
h_+(x,t,\lambda) &= \sum_{n=1}^{N}\lambda^{-n}\partial_x f_{N-n}(x,t)\,, \\
g_-(z,t,\lambda)\exp(-i(\lambda z+\lambda^3 t)) &= 1+\sum_{n=1}^{N}(-\lambda)^{-n}f_{N-n}(z,t)\,.
\end{aligned}
\tag{33}
$$

Therefore, for all λ with $|\lambda| = R > 1$,

$$
|h_+(x,t,\lambda)| \leq \frac{1}{R}\sum_{n=1}^{N}|\partial_x f_{N-n}(x,t)| \tag{34}
$$

and

$$
\sup_{z\in\mathbb{R}}|g_-(z,t,\lambda)\exp(-i(\lambda z+\lambda^3 t))| \leq 1+\sup_{z\in\mathbb{R}}\sum_{n=1}^{N}|f_{N-n}(z,t)|\,. \tag{35}
$$

This latter relation assumes the uniform boundedness of the functions $f_k(z,t)$ in z. Finally, it is clear that for $|\lambda| = R > \sup_n 2\eta_n$

$$
\left|\lambda^{1-2N}\prod_{n=1}^{N}(\lambda^2+4\eta_n^2)\right| \geq R\prod_{n=1}^{N}\left(1-\frac{4\eta_n^2}{R^2}\right). \tag{36}
$$

It follows that for all λ with $|\lambda| = R$ sufficiently large,

$$
|J_U| \leq \frac{K(x,t)}{R}\|\phi\|_1 \tag{37}
$$

where

$$
K(x,t) = \pi\prod_{n=1}^{N}\left(1-\frac{4\eta_n^2}{R^2}\right)^{-1}\cdot\left(\sum_{n=1}^{N}|\partial_x f_{N-n}(x,t)|\right)\cdot\left(1+\sup_{z\in\mathbb{R}}\sum_{n=1}^{N}|f_{N-n}(z,t)|\right). \tag{38}
$$

The bound (37) clearly vanishes as $R \to \infty$. A nearly identical argument shows that the integral

$$J_L \quad := \quad \int_{C_L} \partial_x \left[-\int_{x_{0,L}}^{x} \frac{g_-(z,t,\lambda)\exp(-i(\lambda z+\lambda^3 t))\phi(z)}{\lambda^{1-2N}\prod_{n=1}^{N}(\lambda^2+4\eta_n^2)}\,dz\cdot g_+(x,t,\lambda)\right] d\lambda \tag{39}$$

satisfies the same bound (37) as J_U.

Now we consider integrating the second terms of $\partial_x\psi_U(x,t,\lambda)$ and $\partial_x\psi_L(x,t,\lambda)$, respectively. For brevity, define

$$Y(x,z,t,\lambda) := \frac{g_+(z,t,\lambda)\exp(-i(\lambda z+\lambda^3 t))h_-(x,t,\lambda)}{\lambda^{1-2N}\prod_{n=1}^{N}(\lambda^2+4\eta_n^2)}\,. \tag{40}$$

Note that this can be written as

$$\begin{aligned} Y(x,z,t,\lambda) &= \exp\left(i\lambda(x-z)\right)\frac{\left(1+\sum_{n=1}^{N}\frac{f_{N-n}(z,t)}{\lambda^n}\right)}{\lambda\prod_{n=1}^{N}\left(1+\frac{4\eta_n^2}{\lambda^2}\right)} \\ &\quad\times\ \left(i\lambda\left(1+\sum_{n=1}^{N}\frac{f_{N-n}(x,t)}{(-\lambda)^n}\right)+\sum_{n=1}^{N}\frac{\partial_x f_{N-n}(x,t)}{(-\lambda)^n}\right), \end{aligned} \tag{41}$$

and therefore,

$$Y(x,z,t,\lambda) = i\exp(i\lambda(x-z))\left(1+\Delta(x,z,t,\lambda)\right)\,, \tag{42}$$

where $\Delta(x,z,t,\lambda)=O(\lambda^{-1})$ uniformly in x and z for fixed t. It also follows from additional cancellation that for $z=x$, $\Delta(x,x,t,\lambda)=O(\lambda^{-2})$ uniformly in x. Finally, derivatives of Δ are controlled as well: $\partial_z\Delta(x,z,t,\lambda)=O(\lambda^{-1})$ uniformly. The integral we need to compute for the contribution of $\partial_x\psi_U(x,t,\lambda)$ is

$$\begin{aligned} \int_{C_U}\int_{-\infty}^{x} Y(x,z,t,\lambda)\phi(z)\,dz\,d\lambda = {}& i\int_{C_U}\int_{-\infty}^{x}\exp(i\lambda(x-z))\phi(z)\,dz\,d\lambda \\ &+\ i\int_{C_U}\int_{-\infty}^{x}\Delta(x,z,t,\lambda)\exp(i\lambda(x-z))\phi(z)\,dz\,d\lambda\,. \end{aligned} \tag{43}$$

Note that since the integrand is analytic in the upper half-plane, the first term can be written as

$$i\int_{C_U}\int_{-\infty}^{x}\exp(i\lambda(x-z))\phi(z)\,dz\,d\lambda = -i\int_{-R}^{R}\int_{-\infty}^{x}\exp(i\lambda(x-z))\phi(z)\,dz\,d\lambda\,. \tag{44}$$

In order to control the error term, it is necessary to integrate by parts once:

$$\begin{aligned} i\int_{C_U}\int_{-\infty}^{x}\Delta(x,z,t,\lambda)\exp(i\lambda(x-z))\phi(z)\,dz\,d\lambda &= -\phi(x)\int_{C_U}\frac{\Delta(x,x,t,\lambda)}{\lambda}\,d\lambda \\ +\ \int_{C_U}\int_{-\infty}^{x}\frac{\exp(i\lambda(x-z))}{\lambda}&\partial_z(\Delta(x,z,t,\lambda)\phi(z))\,dz\,d\lambda\,. \end{aligned} \tag{45}$$

The boundary term at $z=-\infty$ vanishes because $\Delta(x,z,t,\lambda)$ is bounded there, ϕ is continuous and integrable, and $\exp(i\lambda(x-z))$ is exponentially small for $\Im(\lambda)>0$. Since the exponential is bounded in magnitude by unity for $\Im(\lambda)>0$ and the contour is of length πR, the above estimates of Δ imply that there exist $K_0(x,t)$, $K_1(x,t)$, and $K_2(x,t)$ all positive, such that

$$\begin{aligned}&\left|\int_{C_U}\int_{-\infty}^{x}\Delta(x,z,t,\lambda)\exp(i\lambda(x-z))\phi(z)\,dz\,d\lambda\right|\\ &\qquad=\frac{K_0(x,t)}{R^2}|\phi(x)|+\frac{K_1(x,t)}{R}\|\phi\|_1+\frac{K_2(x,t)}{R}\|\phi'\|_1\,.\end{aligned}\tag{46}$$

This proves that

$$\lim_{R\to\infty}\int_{C_U}\int_{-\infty}^{x}Y(x,z,t,\lambda)\phi(z)\,dz\,d\lambda=-i\lim_{R\to\infty}\int_{-R}^{R}\int_{-\infty}^{x}\exp(i\lambda(x-z))\phi(z)\,dz\,d\lambda\,.\tag{47}$$

Similar arguments applied to the contribution of $\partial_x\psi_L(x,t,\lambda)$ show that

$$-\lim_{R\to\infty}\int_{C_L}\int_{x}^{\infty}Y(x,z,t,\lambda)\phi(z)\,dz\,d\lambda=-i\lim_{R\to\infty}\int_{-R}^{R}\int_{x}^{\infty}\exp(i\lambda(x-z))\phi(z)\,dz\,d\lambda\,,\tag{48}$$

and therefore,

$$\begin{aligned}&-\frac{1}{2\pi i}\lim_{R\to\infty}\left[\int_{C_U}\partial_x\psi_U(x,t,\lambda)\,d\lambda+\int_{C_L}\partial_x\psi_L(x,t,\lambda)\,d\lambda\right]\\ &\qquad=\frac{1}{2\pi}\lim_{R\to\infty}\int_{-R}^{R}\int_{-\infty}^{\infty}\exp(i\lambda(x-z))\phi(z)\,dz\,d\lambda=\phi(x)\end{aligned}\tag{49}$$

with the last equality following from Fourier inversion. This establishes (31) and the theorem. □

As it stands, the completeness relation given in Theorem 3.1 is not really an expansion of $\phi(x)$ in terms of the functions $h_-(x,t,\lambda)$ because the expansion coefficients themselves depend on x. This is easily remedied by casting the right-hand side of the completeness relation into a more useful form. We do this now.

THEOREM 3.2. *Let $\phi(x)$ be an absolutely continuous function in $L^1(\mathbb{R})$. Let $t\in\mathbb{R}$ be fixed, and choose any $w\in\mathbb{R}\cup\{-\infty,+\infty\}$. Define the mode function*

$$H(x,t,\lambda):=\lambda^N h_-(x,t,\lambda)\,,\tag{50}$$

which is an entire function of λ, and the amplitudes

$$\begin{aligned}b^+(t,\lambda)&:=\int_w^{\infty}\frac{\lambda^N g_+(z,t,\lambda)\exp(-i(\lambda z+\lambda^3 t))}{\lambda\prod_{n=1}^{N}(\lambda^2+4\eta_n^2)}\phi(z)\,dz\,,\\ b^-(t,\lambda)&:=\int_{-\infty}^{w}\frac{\lambda^N g_+(z,t,\lambda)\exp(-i(\lambda z+\lambda^3 t))}{\lambda\prod_{n=1}^{N}(\lambda^2+4\eta_n^2)}\phi(z)\,dz\,,\end{aligned}\tag{51}$$

and set $b(t,\lambda) := b^+(t,\lambda) + b^-(t,\lambda)$. The amplitudes have simple poles at $\lambda = 0$ and $\lambda = \pm 2i\eta_n$ for $n = 1,\ldots,N$. Finally, set

$$b_0(t) := \frac{1}{2} \operatorname*{Res}_{\lambda=0} \left(b^+(t,\lambda) - b^-(t,\lambda)\right) \qquad \textit{and} \qquad b_n^{\pm}(t) := \mp \operatorname*{Res}_{\lambda=\pm 2i\eta_n} b^{\mp}(t,\lambda)\,. \tag{52}$$

Then we have the expansion

$$\begin{aligned}\phi(x) = {} & \lim_{R\to\infty} \frac{1}{2\pi i} \mathrm{P.V.} \int_{-R}^{R} b(t,\lambda) H(x,t,\lambda)\, d\lambda \\ & + b_0(t) H(x,t,0) + \sum_{n=1}^{N} \left[b_n^-(t) H(x,t,-2i\eta_n) + b_n^+(t) H(x,t,2i\eta_n)\right]\,.\end{aligned} \tag{53}$$

Remark 1. Since w is now fixed and not a function of x, this expansion (53) is a true completeness relation, expressing an arbitrary given function $\phi(x)$ as a sum of known functions $H(x,t,\lambda)$. From the exact formulas (50), (15), and (8), it is clear that the part of the expansion (53) represented by the singular integral is Fourier-like, with the corresponding components of the solution, $H(x,t,\lambda)$ for $\lambda \in \mathbb{R}$ being bounded oscillatory functions tending to complex exponentials for large x. On the other hand, the discrete contributions to the solution represent bound states. The $2N+1$ bound state terms in (53) are not linearly independent. From the fact that at the eigenvalues $\pm 2i\eta_n$ the functions $g_-(x,t,\lambda)$ are all linear combinations of the same N functions $f_0(x,t),\ldots,f_{N-1}(x,t)$, it is clear that only N of the bound states are linearly independent. These facts are easiest to see when one takes w to ∞ or $-\infty$. Then, half of the contributions from the eigenvalues disappear, and it remains only to express the bound state at zero, $H(x,t,0)$, in terms of $H(x,t,\pm 2i\eta_n)$. This can be done directly. From the exact formulas (50), (15), and (8), we see that

$$H(x,t,0) = (-1)^N \partial_x f_0(x,t)\,, \tag{54}$$

and making use of the relations (9) satisfied by f_+ at the eigenvalues,

$$H(x,t,2i\eta_n) = (-1)^{n+1} \exp(-2\eta_n\alpha_n) \sum_{p=0}^{N-1} (2i\eta_n)^p \partial_x f_p(x,t)\,. \tag{55}$$

Expressing $H(x,t,0)$ in terms of $H(x,t,2i\eta_n)$ is therefore a polynomial interpolation problem. Introduce the polynomial

$$P(\lambda) = \sum_{p=0}^{N-1} \partial_x f_p(x,t)\lambda^p\,. \tag{56}$$

Given isolated values of this polynomial

$$P(2i\eta_n) = (-1)^{n+1} \exp(2\eta_n\alpha_n) H(x,t,2i\eta_n) \qquad \text{for} \qquad n = 1,\ldots,N\,, \tag{57}$$

we are to find $P(0)$ and thus $H(x,t,0) = (-1)^N \partial_x f_0(x,t) = (-1)^N P(0)$. Expressing $P(\lambda)$ explicitly in terms of Lagrange polynomials gives

$$P(\lambda) = \sum_{n=1}^{N} (-1)^{n+1} \exp(2\eta_n\alpha_n) H(x,t,2i\eta_n) \prod_{k\neq n} \frac{\lambda - 2i\eta_k}{2i\eta_n - 2i\eta_k}\,, \tag{58}$$

and therefore

$$H(x,t,0)=\sum_{n=1}^{N}\left[(-1)^{n}\exp(2\eta_n\alpha_n)\prod_{k\neq n}\frac{\eta_k}{\eta_n-\eta_k}\right]H(x,t,2i\eta_n)\,. \tag{59}$$

There is, indeed, a similar expression for $H(x,t,0)$ in terms of $H(x,t,-2i\eta_n)$. □

Remark 2. An important distinction between the completeness relation stated in Theorem 3.2 and that found by Sachs [S83] for derivatives of squared Schrödinger eigenfunctions is in the nature of the singularity at $\lambda=0$. Sachs shows that in the expansion of ϕ in terms of derivatives of squared eigenfunctions, there is an apparent singularity at $\lambda=0$ that is in fact removable. On the other hand, the integral in Theorem 3.2 is essentially singular and the residue contribution of $b_0(t)$ is nonzero. □

Proof of Theorem 3.2. We first establish that in the formulas (29) for $\partial_x\psi_U(x,t,\lambda)$ and (30) for $\partial_x\psi_L(x,t,\lambda)$ we may replace x in the limits of integration by any other value without changing the result of the theorem. That is, we will now show that the integral

$$\int_{C_U}\int_{-\infty}^{w}Y(x,z,t,\lambda)\phi(z)\,dz\,d\lambda-\int_{C_L}\int_{w}^{\infty}Y(x,z,t,\lambda)\phi(z)\,dz\,d\lambda\,, \tag{60}$$

which we have already seen converges as R tends to infinity to $-2\pi i\phi(x)$ in the case that $w=x$, is in fact independent of w. Holding R fixed and differentiating with respect to w, we must show that for sufficiently large R,

$$\phi(w)\oint_{|\lambda|=R}Y(x,w,t,\lambda)\,d\lambda\equiv 0 \tag{61}$$

identically in x, w, and t. Being as the integrand is meromorphic in the finite λ plane, we can evaluate the integral by residues. There are simple poles at $\lambda=0$ and $\lambda=\pm 2i\eta_n$ for $n=1,\ldots,N$. Using the linear relations (9) satisfied by $f_\pm$ at the eigenvalues $\lambda=2i\eta_n$, we find

$$\begin{aligned}\operatorname*{Res}_{\lambda=2i\eta_k}\;Y(x,w,t,\lambda)&=\sum_{p=0}^{N-1}\frac{(2i\eta_k)^{N+p-1}}{D_k}\partial_x f_p(x,t)\\&\quad+\sum_{p,q=0}^{N-1}(-1)^{N-q}\frac{(2i\eta_k)^{p+q-1}}{D_k}f_q(w,t)\partial_x f_p(x,t)\,,\\\operatorname*{Res}_{\lambda=-2i\eta_k}\;Y(x,w,t,\lambda)&=\sum_{p=0}^{N-1}(-1)^{N-p}\frac{(2i\eta_k)^{N+p-1}}{D_k}\partial_x f_p(x,t)\\&\quad+\sum_{p,q=0}^{N-1}(-1)^{N-p}\frac{(2i\eta_k)^{p+q-1}}{D_k}f_q(w,t)\partial_x f_p(x,t)\,,\end{aligned} \tag{62}$$

where

$$D_k:=\prod_{n\neq k}(2i\eta_k-2i\eta_n)\prod_{n=1}^{N}(2i\eta_k+2i\eta_n)\,. \tag{63}$$

Similarly, for the residue at zero,

$$\operatorname*{Res}_{\lambda=0} \; Y(x,w,t,\lambda) = (-1)^N \frac{f_0(w,t)\partial_x f_0(x,t)}{\displaystyle\prod_{n=1}^{N} 4\eta_n^2}\,. \tag{64}$$

Adding all the residues and collecting coefficients of the terms $\partial_x f_p(x,t)$ and $f_q(w,t)$ $\partial_x f_p(x,t)$, we find that the sum of the residues will be zero if

$$I_p := \sum_{k=1}^{N} \frac{(2i\eta_k)^p}{D_k} = 0 \tag{65}$$

for all odd $p = 1, 3, 5, \ldots, 2N-3$, and if

$$\frac{1}{\displaystyle\prod_{n=1}^{N} 4\eta_n^2} + 2\sum_{k=1}^{N} \frac{1}{2i\eta_k D_k} = 0\,. \tag{66}$$

These expressions are themselves sums of residues of *meromorphic* differentials. Thus, by inspection, one finds that for $p = 1, 3, 5, \ldots, 2N-3$,

$$I_p = \frac{1}{2\pi i} \oint_C \frac{\lambda^p \, d\lambda}{\displaystyle\prod_{n=1}^{N} (\lambda^2 + 4\eta_n^2)}\,, \tag{67}$$

where C is any simple counterclockwise oriented contour that encircles the points $\lambda = 2i\eta_n$ for $n = 1, \ldots, N$ (but without enclosing the conjugate eigenvalues or $\lambda = 0$). With p bounded by $2N-3$, the path of integration can be blown out to infinity in the upper half-plane and then brought down to the real axis so that

$$I_p = \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{\lambda^p \, d\lambda}{\displaystyle\prod_{n=1}^{N} (\lambda^2 + 4\eta_n^2)} = 0 \tag{68}$$

with the last equality following from the oddness of the integrand for odd p. Finally, consider the integral I_{-1} defined by

$$I_{-1} := \frac{1}{2\pi i} \oint_C \frac{d\lambda}{\lambda \displaystyle\prod_{n=1}^{N} (\lambda^2 + 4\eta_n^2)}\,. \tag{69}$$

Evaluating the residues inside C, we find

$$I_{-1} = \sum_{k=1}^{N} \frac{1}{2i\eta_k \displaystyle\prod_{n\neq k} (2i\eta_k - 2i\eta_n) \prod_{n=1}^{N} (2i\eta_k + 2i\eta_n)}\,. \tag{70}$$

On the other hand, we can again blow the contour C out to infinity in the upper half-plane and bring it down to the real axis. This time, there is a singularity at $\lambda = 0$, so the Plemelj formula must be used. We find

$$I_{-1} = -\frac{1}{2} \cdot \frac{1}{\displaystyle\prod_{n=1}^{N} 4\eta_n^2} + \frac{1}{2\pi i}\text{P. V.} \int_{-\infty}^{\infty} \frac{d\lambda}{\lambda \displaystyle\prod_{n=1}^{N}(\lambda^2 + 4\eta_n^2)} . \tag{71}$$

Once again, by oddness, the principal value integral vanishes identically, and then combining this result with the previous expression, we obtain the required vanishing.

This shows that for any w,

$$\phi(x) = -\frac{1}{2\pi i} \lim_{R\to\infty} \left[\int_{C_U} \int_{-\infty}^{w} Y(x,z,t,\lambda)\phi(z)\,dz\,d\lambda - \int_{C_L} \int_{w}^{\infty} Y(x,z,t,\lambda)\phi(z)\,dz\,d\lambda\right] . \tag{72}$$

Establishing (53) and therefore the theorem now amounts to using the residue theorem once again to deform the integration paths C_U and C_L in (72) to the real axis. One finds discrete contributions at the poles $\lambda = \pm 2i\eta_n$, and then applying the Plemelj formula to contract the contour to the real axis in the neighborhood of $\lambda = 0$ gives a discrete contribution proportional to $H(x,t,0)$ and the principal value regularization of the singular integral over the continuous spectrum. □

4. Solution of the initial value problem for $\boldsymbol{\kappa = 1/2}$. It is easy to see that when $A(x,t)$ is an N-soliton solution of KdV (1), one can use the completeness relation to solve the initial value problem

$$\partial_t B + \partial_x \left[\frac{1}{2}AB + \partial_x^2 B\right] = 0\,, \qquad B(x,0) = \phi(x)\,. \tag{73}$$

Setting $t = 0$, and picking a convenient value of w, say, $w = +\infty$, one computes the amplitudes (51) and discrete coefficients (52). Then, because the function $H(x,t,\lambda)$ satisfies (2) for $\kappa = 1/2$ and for each complex λ, the expression

$$\begin{aligned} B(x,t) := \quad & \lim_{R\to\infty} \frac{1}{2\pi i}\text{P. V.} \int_{-R}^{R} b(0,\lambda)H(x,t,\lambda)\,d\lambda \\ & + b_0(0)H(x,t,0) + \sum_{n=1}^{N} \left[b_n^-(0)H(x,t,-2i\eta_n) + b_n^+(0)H(x,t,2i\eta_n)\right] , \end{aligned} \tag{74}$$

provides the solution of the initial value problem (73), generally in the sense of distributions. That is, $B(x,0) = \phi(x)$ by Theorem 3.2, and for each test function $\varphi(x,t)$ that is differentiable in t and three times differentiable in x and has compact support in $(x,t) \in \mathbb{R} \times \mathbb{R}_+$, one shows by exchanging the order of integration that

$$\int_0^{\infty} \int_{-\infty}^{\infty} \left[\partial_t\varphi(x,t) + \frac{1}{2}A(x,t)\partial_x\varphi(x,t) + \partial_x^3\varphi(x,t)\right] B(x,t)\,dx\,dt = 0\,. \tag{75}$$

The solution will be classical in as much as it is possible to differentiate with respect to x and t under the integral sign in the solution formula (74). This requires additional smoothness and decay assumptions on the initial data $\phi(x)$ that we do not consider here.

5. Scattering of bound states for $\kappa = 1/2$. Of particular interest in applications is the N-dimensional (recall that $A(x,t)$ is an N-soliton solution of KdV) subspace of solutions of (2) for $\kappa = 1/2$ consisting of bound states. This subspace represents linear waves that are trapped by the solitons of the potential $A(x,t)$. For large $|t|$, these bound state solutions are all confined to the trajectories of the solitons. Therefore, it follows that each bound state $B(x,t)$ has two asymptotic representations:

$$B(x,t) \sim \sum_{n=1}^{N} \beta_n^{\pm} A_n^{\pm}(x,t)\,, \qquad t \to \pm\infty\,, \tag{76}$$

for some constants $\beta_n^{\pm}$ depending on $B(x,t)$, where $A_n^{\pm}(x,t)$ are defined by (11). Since there are exactly N linearly independent bound states, it follows that the constants β_n^{+} are completely determined from the constants β_n^{-}. In particular, there exists an invertible $N \times N$ matrix $\mathbf{T}$ with entries depending only on the data specifying the N-soliton solution $A(x,t)$, such that

$$\beta_j^{+} = \sum_{k=1}^{N} T_{jk}\beta_k^{-}\,. \tag{77}$$

The matrix $\mathbf{T}$ is called the *bound state scattering matrix.* In this section, we compute the scattering matrix explicitly and show that its elements only depend on the soliton eigenvalues $\eta_1, \ldots, \eta_N$.

If $A(x,t)$ is an N-soliton solution of KdV (1), then a family of solutions to (2) for $\kappa = 1/2$, parametrized by complex λ, is given by

$$h_+(x,t,\lambda) = \frac{A(x,t)}{6i\lambda} + \sum_{n=0}^{N-2} \lambda^{n-N}\partial_x f_n(x,t)\,. \tag{78}$$

We want to analyze these solutions in the limit of large $|t|$, in a frame of reference traveling with constant velocity c.

The first step is to see how the coefficients $f_n(x,t)$ behave for large $|t|$. Let $\chi = x - ct$ be fixed as $\tau = t$ goes to either $+\infty$ or $-\infty$. Begin by taking $\eta_m^2 < 4c < \eta_{m-1}^2$ to see how the coefficients behave in between the solitons. In the limit $\tau \to +\infty$, the equations (9) imply that

$$\begin{aligned} 1 \quad &+ \quad \sum_{k=0}^{N-1} (-2i\eta_n)^{k-N} f_k \quad \to \quad 0\,, \quad n = 1, \ldots, m-1\,, \\ 1 \quad &+ \quad \sum_{k=0}^{N-1} (2i\eta_n)^{k-N} f_k \quad \to \quad 0\,, \quad n = m, \ldots, N\,. \end{aligned} \tag{79}$$

This is an invertible Vandermonde system for the coefficients f_k, so that as $\tau \to +\infty$, the f_k all become constants, independent of χ and τ. Thus, $\partial_x f_k(x,t)$ vanishes between the solitons for all k. The analogous result holds as $\tau \to -\infty$. This shows that the solutions of (2) for $\kappa = 1/2$ described by the formula (78) are asymptotically confined to the individual frames of reference of the moving solitons in the potential field $A(x,t)$.

Now set $c = 4\eta_m^2$ to go into the moving frame of reference of one of the solitons. Taking the limit $\tau \to +\infty$ yields

$$
\begin{aligned}
1 \quad + \quad & \sum_{k=0}^{N-1} (-2i\eta_n)^{k-N} f_k \quad \to \quad 0 \,, \quad n = 1, \ldots, m-1 \,, \\
1 \quad + \quad & \sum_{k=0}^{N-1} (2i\eta_n)^{k-N} f_k \quad \to \quad 0 \,, \quad n = m+1, \ldots, N \,.
\end{aligned}
\tag{80}
$$

This is a system of $N-1$ equations in N unknowns, so it can be used to asymptotically eliminate $\partial_\chi f_0$ through $\partial_\chi f_{N-2}$ in favor of $\partial_\chi f_{N-1}$, which we know is proportional to the N-soliton solution of KdV, $A(x,t)$. Thus, as $\tau \to +\infty$, with $c = 4\eta_m^2$,

$$
\partial_\chi f_k = Q_{mk} \partial_\chi f_{N-1} = \frac{1}{6i} Q_{mk} A
\tag{81}
$$

for $k = 0, \ldots, N-2$, where the numbers Q_{mk} are the unique solution of the inhomogeneous system of linear algebraic equations

$$
\begin{aligned}
(-2i\eta_n)^{-1} \quad + \quad & \sum_{k=0}^{N-2} (-2i\eta_n)^{k-N} Q_{mk} \quad = \quad 0 \,, \quad n = 1, \ldots, m-1 \,, \\
(2i\eta_n)^{-1} \quad + \quad & \sum_{k=0}^{N-2} (2i\eta_n)^{k-N} Q_{mk} \quad = \quad 0 \,, \quad n = m+1, \ldots, N \,.
\end{aligned}
\tag{82}
$$

One can similarly show that as $\tau \to -\infty$, with $c = 4\eta_m^2$,

$$
\partial_\chi f_k = Q_{mk}^* \partial_\chi f_{N-1} = \frac{1}{6i} Q_{mk}^* A
\tag{83}
$$

for $k = 0, \ldots, N-2$, where the star denotes complex conjugation.

Now consider particular solutions $B_j(x,t)$ of (2) for $\kappa = 1/2$ obtained as linear combinations of N others expressed by the formula (78) evaluated on the N soliton eigenvalues. The formula for $B_j(x,t)$ is

$$
B_j(x,t) = \sum_{k=1}^{N} F_{jk} h_+(x,t,2i\eta_k) = \sum_{k=1}^{N} F_{jk} \left[-\frac{A(x,t)}{12\eta_k} + \sum_{n=0}^{N-2} (2i\eta_k)^{n-N} \partial_x f_n(x,t) \right] \,,
\tag{84}
$$

where $\mathbf{F} = \{F_{jk}\}$ is a matrix of arbitrary constants. From the asymptotics of $f_n(x,t)$, we have as $\tau \to -\infty$ with $c = 4\eta_m^2$

$$
B_j \to A \sum_{k=1}^{N} F_{jk} G_{km}^- \,, \qquad \text{where} \qquad G_{km}^- := -\frac{1}{12\eta_k} + \frac{1}{6i} \sum_{n=0}^{N-2} (2i\eta_k)^{n-N} Q_{mn}^* \,.
\tag{85}
$$

So, with the choice that the matrix $\{F_{jk}\}$ is the inverse of the matrix $\mathbf{G}^- = \{G_{km}^-\}$, the particular solution $B_j(x,t)$ of (2) for $\kappa = 1/2$ will be completely confined as $t \to -\infty$ to the frame of reference moving with speed $c = 4\eta_j^2$, where it will be locally indistinguishable from the solution $A(x,t)$ of KdV. Let us now determine how $B_j(x,t)$

will behave in the various soliton frames as $t \to +\infty$. Passing to the limit of $\tau \to +\infty$ in the frame with velocity $c = 4\eta_m^2$ gives

$$(86)\quad B_j \to A \sum_{k=1}^{N} F_{jk} G_{km}^{+}\,, \qquad \text{where} \qquad G_{km}^{+} := -\frac{1}{12\eta_k} + \frac{1}{6i}\sum_{n=0}^{N-2} (2i\eta_k)^{n-N} Q_{mn}\,.$$

These asymptotics give us a formula for the bound state scattering matrix:

$$(87) \qquad \mathbf{T} := \left[(\mathbf{G}^{-})^{-1}\mathbf{G}^{+}\right]^{T}\,.$$

It is clear that the elements of $\mathbf{T}$ depend only on the N soliton eigenvalues $\eta_1, \dots, \eta_N$. There is no dependence on the soliton phase variables $\alpha_1, \dots, \alpha_N$. Therefore, the asymptotic scattering properties of linear waves in (2) with $\kappa = 1/2$ are insensitive to phase shifts among the solitons in the potential $A(x,t)$. As a concrete example of the scattering matrix, we compute it explicitly for $N = 2$ for arbitrary $\eta_1 > \eta_2 > 0$:

$$(88) \qquad \mathbf{T} = \frac{1}{\eta_1^2 - \eta_2^2}\begin{bmatrix} (\eta_1 - \eta_2)^2 & 2\eta_2(\eta_1 - \eta_2) \\ 2\eta_1(\eta_1 - \eta_2) & -(\eta_1 - \eta_2)^2 \end{bmatrix}\,.$$

The fact that T_{22} is negative means that it is possible for the interactions of the solitons in $A(x,t)$ to convert trapped linear waves of elevation into waves of depression, and vice-versa.

6. General values of $\boldsymbol{\kappa}$. We expect that for most values of κ, the linear waves satisfying (2) will not be permanently trapped by solitons present in the potential $A(x,t)$. This is suggested by considering the simplest case, namely, taking $A(x,t)$ to be the one-soliton solution of KdV (1). The soliton travels with velocity $c = 4\eta^2$ so that $A = -V(\chi)$ with $\chi = x - ct - \alpha$. Corresponding traveling wave solutions $B(\chi)$ of the linear problem that propagate with the same velocity and decay as $\chi \to \pm\infty$ satisfy

$$(89) \qquad -B''(\chi) + \kappa V(\chi)B(\chi) = -cB(\chi)\,.$$

Since c is fixed, we can view this as an eigenvalue equation with κ as the eigenvalue. We therefore expect that only isolated values of κ will admit nontrivial decaying solutions $B(\chi)$. We have already seen that $\kappa = 1/2$ and $\kappa = 1$ are indeed eigenvalues. For $\kappa = 1/2$ the eigenfunction $B(\chi)$ is an even function of χ, while for $\kappa = 1$ the eigenfunction $B(\chi)$ is odd in χ. Since eigenfunctions of (89) must be nondegenerate and therefore have either odd or even parity in χ, there cannot exist a nontrivial bound state eigenfunction of (89) for all $\kappa \in [1/2, 1]$ because the eigenfunction would have to change parity from one endpoint to the other. Therefore, at least one value of $\kappa \in [1/2, 1]$ is not an eigenvalue. For such κ, there is no bound state traveling wave solution of (2) that is trapped in the soliton trajectory.

We can be more precise about this phenomenon. The left-hand side of (89) can also be viewed as a Schrödinger operator $L(\kappa)$ depending on a coupling constant κ, and the condition for wave trapping by solitons is simply that $-c \in \Sigma_{\mathrm{p}}(L(\kappa))$, where Σ_{p} denotes the point spectrum. The number of discrete eigenvalues is a nondecreasing function of $\kappa > 0$, corresponding to the deepening of the potential well. There is an infinite unbounded sequence of *cutoff* values κ_n^{cut} of κ at which the number of eigenvalues changes by one, and the new eigenvalue is born from the continuum. Each eigenvalue, once born, is distinct and is a decreasing function of κ. From these arguments, it follows that there exists an infinite unbounded sequence of *bifurcation* values κ_n^{bif} of κ at which one eigenvalue crosses the level $E = -c$, and a bound state traveling wave solution of (2) exists.

It is easy to find the bifurcation points because the hyperbolic secant squared potential is so well understood. The potential $\kappa V(\chi)$ is exactly reflectionless for $12\kappa = n(n+1)$ for $n = 1, 2, 3, \ldots$. The corresponding energy levels are $E_{n,k} = -k^2\eta^2$ for $k = 1, \ldots, n$. Therefore, for $n > 1$ in this sequence, there is always one eigenvalue that is exactly equal to $-c = -4\eta^2$. The corresponding eigenstate is always the $(n-1)$st state and therefore has $n-2$ zeros. It follows that the bifurcation points are $\kappa = \kappa_n^{\mathrm{bif}} = (n+1)(n+2)/12$ for $n = 1, 2, 3, \ldots$.

The fact that some linear waves may be permanently trapped by isolated solitons at a bifurcation point $\kappa = \kappa_n^{\mathrm{bif}}$ does not necessarily imply that there will be no losses to radiation when solitons in the field $A(x,t)$ interact with one another. Such a lossless interaction might suggest the "integrability" of the linear equation (2). We have indeed seen that this is the case for the first two bifurcation points, $\kappa = \kappa_1^{\mathrm{bif}} = 1/2$ and $\kappa = \kappa_2^{\mathrm{bif}} = 1$, but it is by no means clear that the trend continues for higher-order bifurcation points. For the rest of this section, we therefore restrict attention to the case $N = 1$, that is, we take the nonlinear field $A(x,t)$ to be a one-soliton solution of KdV (1).

Using (3) and the change of variables $x' = \eta(x - 4\eta^2 t - \alpha)$ and $t' = \eta^3 t$, (2) becomes, after dropping primes,

$$\partial_t B + \partial_x[-4B + 12\kappa\,\mathrm{sech}^2(x)B + \partial_x^2 B] = 0\,. \tag{90}$$

This equation is of course solved by separation of variables. We seek separated solutions $B(x,t) = b_\sigma(x)\exp(\sigma t)$ and obtain the third-order eigenvalue problem

$$[4b_\sigma(x) - 12\kappa\,\mathrm{sech}^2(x)b_\sigma(x) - b_\sigma''(x)]' = \sigma b_\sigma(x)\,, \tag{91}$$

where the prime denotes differentiation with respect to x. In this context, what we have been calling "trapped linear waves" correspond to bound-state eigenfunctions of (91) with $\sigma = 0$. Such solutions have finite mass and energy and are stationary in the moving frame of reference of the soliton $A(x,t)$. As we know, such eigenfunctions with $\sigma = 0$ exist only at the bifurcation values of $\kappa = \kappa_n^{\mathrm{bif}}$. However, it is clear that for general values of κ there are other possibilities. There may be eigenvalues σ that are purely imaginary, giving rise to oscillating modes that travel in the soliton frame. More generally, if an eigenvalue has a nonzero real part for some κ, then there will be a mode that is either amplified or exponentially damped as it propagates with the soliton.

The eigenvalue problem (91) has two simple symmetries. Whenever $b_\sigma(x)$ is an eigenfunction with eigenvalue σ, then $b_\sigma(-x)$ is an eigenfunction with eigenvalue $-\sigma$ and $b_\sigma(x)^*$ is an eigenfunction with eigenvalue σ^*. Therefore, the eigenvalues either come in purely real pairs $(|\sigma|, -|\sigma|)$, purely imaginary pairs $(i|\sigma|, -i|\sigma|)$, or in complex quartets $(\sigma, -\sigma, \sigma^*, -\sigma^*)$. These symmetries indicate the distinguished role of $\sigma = 0$ as a point that if it appears in the spectrum for some κ can signal a bifurcation in the number of eigenvalues. This explains our terminology and notation for the values $\kappa = \kappa_n^{\mathrm{bif}}$.

Most points on the imaginary σ axis correspond to continuous spectrum. This can be seen by the following argument. Let κ be fixed. Suppose $\sigma = i\omega$ with $\omega \in \mathbb{R}$. For large $|x|$, the solutions of (91) have the form of linear combinations of $\exp(ik_{\omega,j}x)$ where $k = k_{\omega,j}$ are the three roots of $k^3 + 4k - \omega = 0$. Exactly one of these roots, say, $k = k_{\omega,0}$, is real, while the other two form a complex-conjugate pair. If we seek a generalized eigenfunction normalized to $\exp(ik_{\omega,0}x)$ as $x \to -\infty$ through a "shooting" method, we have three complex constants to exploit: the coefficient of the decaying mode for large negative x, and the coefficients of the decaying mode for large positive x and the finite amplitude contribution for large positive x. Matching the values of $b_{i\omega}(x)$, $b'_{i\omega}(x)$, and $b''_{i\omega}(x)$ at $x = 0$ gives three complex equations in three complex unknowns. If this system of equations is solvable at all, one expects it to be solvable for almost all real ω, yielding a generalized eigenfunction. For exceptional values of ω where there is not a generalized eigenfunction, there will be a genuine bound-state eigenfunction since the spectrum is a closed set.

We have used a numerical Fourier-based collocation (pseudospectral) method to find the discrete eigenvalues of (91) over a range of values of the coupling constant κ. Essentially this involves approximating the continuous function $b_\sigma(x)$ by a periodic discrete series. Then the derivative ∂_x can be approximated to exponential accuracy by a derivative matrix $\mathbf{D}$ for which an explicit formula is given in [CHQZ88]. This is then used to construct a discrete approximation to the operator on the left-hand side of (91). Standard techniques can then be used to obtain the eigenvalues and eigenvectors of this matrix. The corresponding eigenfunctions always decay exponentially, but sometimes they decay *very* slowly for large x of one or the other sign—luckily not both. To obtain accurate results it was necessary in these cases to change the dependent variable by multiplying by an appropriate exponential function of x to enhance the decay on the slowly decaying side without changing decay into growth on the other side. Our results over the range $0 < \kappa < 5$ are shown in Figure 6.1. The bifurcation values κ_n^{bif} appear to be of two different types. If n is odd, then when κ increases through the value $\kappa = \kappa_n^{\mathrm{bif}}$, a new pair of real eigenvalues is born from the origin $\sigma = 0$. As κ is further increased, the pair of eigenvalues moves at first away from the imaginary axis and then changes direction and contracts toward the origin. When κ increases through the even bifurcation value $\kappa = \kappa_{n+1}^{\mathrm{bif}}$, the pair enters the origin and re-emerges as a complex eigenvalue quartet. Further increasing the value of κ causes the quartet of eigenvalues to move through a maximum in the magnitude of the real part and then toward the imaginary axis with the magnitude of the real part decreasing to zero while the magnitude of the imaginary part increases without bound. It does not appear that the quartet of eigenvalues ever disappears into the continuous spectrum, although it comes arbitrarily close as κ increases. This scenario is repeated again and again as κ increases through each odd bifurcation value. Representative eigenfunctions are plotted in Figure 6.2. Here, one can see that when the eigenvalue σ has a nonzero imaginary part, the decay of the eigenfunction can be quite slow on the "downstream" side of the soliton. As remarked above, this effect is compensated for in our numerics by working with a modified eigenfunction.

There are no discrete eigenvalues at all for $\kappa < 1/2$ (and in particular for $\kappa < 0$), and for all κ satisfying $1/2 < \kappa < 1$ and $\kappa > 1$, there is always at least one eigenvalue with a nonzero real part, which corresponds to an exponentially growing eigenfunction and therefore instability. The values $\kappa = 1/2$ and $\kappa = 1$ are distinguished as the only values for which there exist discrete eigenvalues and at the same time all eigenvalues have zero real parts, so the system is neutrally stable. For all other values of κ, either there are no discrete eigenvalues at all in which case all initial conditions for (2) disperse away algebraically in time, or there are discrete eigenvalues with positive real parts in which case the linear waves are amplified by the soliton in the field $A(x, t)$.

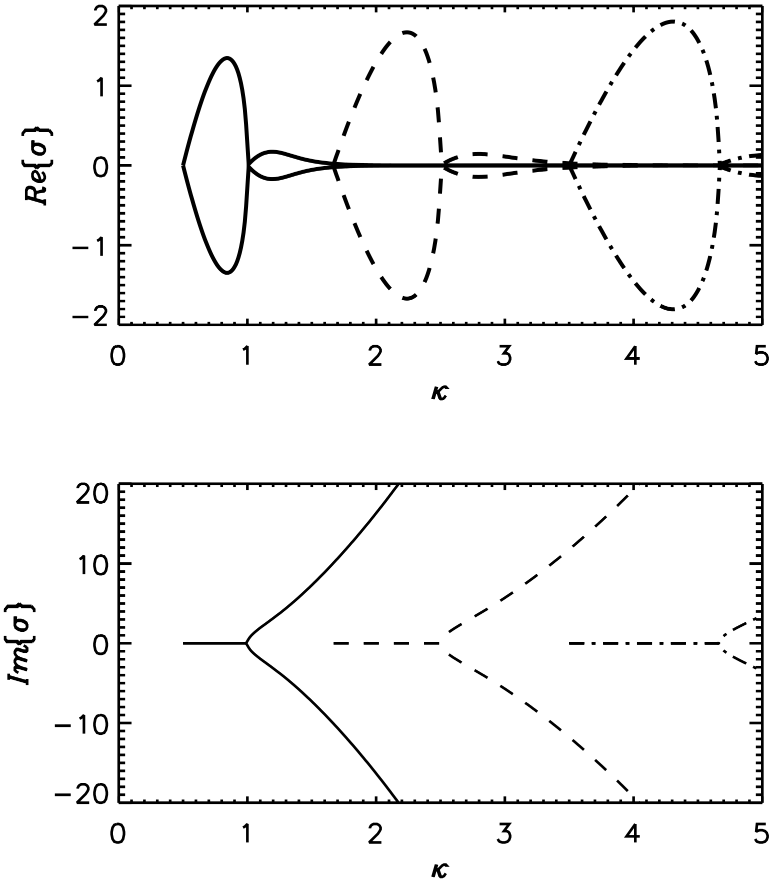


FIG. 6.1. *Real (above) and imaginary (below) parts of the discrete eigenvalues σ for the eigenvalue problem* (91) *as a function of the parameter κ. Different eigenvalue branches are displayed with different styles of lines (solid, dashed, etc.).*

7. Conclusion. The coupled system consisting of the KdV equation (1) and the linear equation (2) is integrable for two distinct values of the coupling parameter κ. The integrable case of $\kappa = 1$ has been studied by other authors [GGKM74, S83]. In this paper, we have given new results for the other integrable case, namely, $\kappa = 1/2$. In particular, we have shown how to construct the general solution of (2) for $\kappa = 1/2$ when the nonlinear field $A(x,t)$ is an N-soliton solution of the KdV equation. This general solution is represented in terms of a number of bound states (equal to the number N of solitons in the field $A(x,t)$) and a continuum superposition of radiative states given by a singular integral. With the help of numerical computations, we have placed the integrable cases in context by examining the behavior of the linear equation (2) for general values of κ, when $A(x,t)$ is a one-soliton solution of KdV. These calculations show that the linear equation (2) behaves as an unstable dynamical system for most positive κ. The integrable value of $\kappa = 1$, for which the equation (2) is the linearized KdV equation, is an isolated stable point, since a small change of either sign in the value of κ will lead to the presence of exponentially growing modes. The other integrable value of $\kappa = 1/2$ represents the boundary between a

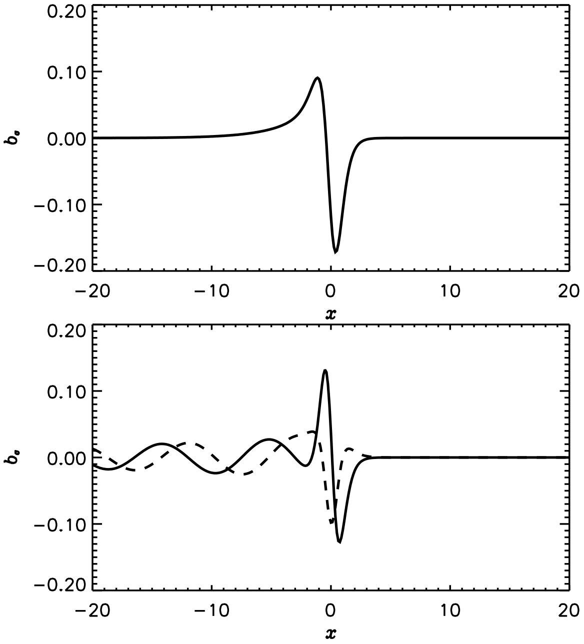


FIG. 6.2. *Above: the real-valued eigenfunction corresponding to the eigenvalue σ with positive real part for $\kappa = 0.85$. Below: the complex-valued eigenfunction corresponding to the eigenvalue σ in the first quadrant for $\kappa = 1.2$. The solid curve is the real part and the dashed curve is the imaginary part.*

stable system without any bound states for $\kappa < 1/2$ and an exponentially unstable system for $\kappa > 1/2$.

In physical applications of the coupled system (1) and (2) as discussed in the appendix, the presence of instabilities indicates that more terms need to be included in the model. However, in the stable cases the model is indeed expected to be physically meaningful. And in this regard, the two integrable cases can provide useful starting points for perturbation theory.

As a final remark, let us indicate the kind of calculations that are possible for the coupled system (1) and (2) for $\kappa = 1/2$ with the aid of the completeness relation. For a family of relevant initial data for the linear equation, one can explicitly compute the projection onto the bound states and consequently determine the long time behavior of the corresponding solution of (2). Also, the long time behavior of the dispersive part of the solution can be computed from the explicit representation of this component of the solution as a singular integral. We leave such applications of the completeness relation for further investigations.

Appendix A. Some applications. It is useful to keep in mind some applications in which the coupled system (1) and (2) might arise. In fact, such equations appear in the modeling of coupling of acoustic phonons in long polymer molecules. Many organic polymers (e.g., DNA and α-helix proteins like acetanelide) may be considered from the mechanical point of view as long chains of nearly identical masses. This "backbone" of the molecule supports a longitudinal vibrational mode in which the masses are all moving in tandem with zero frequency (i.e., simple translation) in

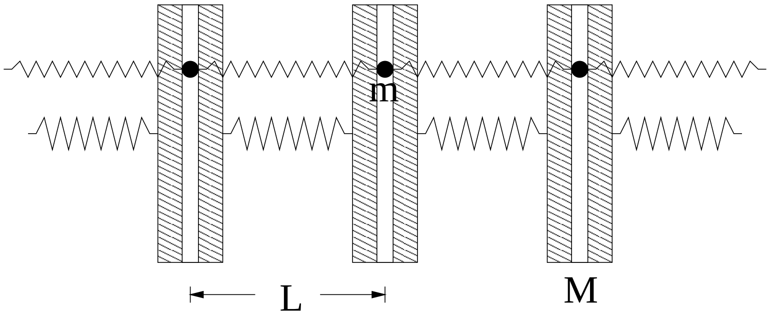


FIG. A.1. *The equilibrium configuration of the mechanical model.*

the long-wave limit; the associated quanta are called acoustic phonons. In the presence of intrinsic weak nonlinearity, the KdV equation describes these vibrations in the long-wave limit. Usually, the masses making up the chain contain internal degrees of freedom (e.g., the "breathing" modes of base-pairs in DNA, and the so-called amide I exciton modes of the C=O bond in each peptide group of an α-helix protein). The coupling of these internal degrees of freedom to the motion of the backbone leads to a variety of interesting dynamical models (e.g., the discrete sine-Gordon equation for DNA and the discrete nonlinear Schrödinger equation for α-helix proteins).

We may consider a situation in which the internal degrees of freedom are themselves acoustic phonons associated with transverse vibrational modes. This can be visualized with the help of a concrete mechanical model, whose equilibrium configuration is shown in Figure A.1. The backbone is made of heavy masses M connected by stiff springs. Mounted on each heavy mass is a transversely-oriented frictionless track in which rides a small mass m. The mass M is assumed to include the mass of the track and small mass m. The small masses are themselves connected by weaker springs. Assigning longitudinal displacements u_n to the large masses M and transverse displacements v_n to the small masses m in the frictionless tracks, the Hamiltonian of the mechanical model is

$$H = \sum_n \left[\frac{1}{2}M\dot{u}_n^2 + \frac{1}{2}m\dot{v}_n^2 + W(L+u_{n+1}-u_n)\right.$$

$$\left. +V\left(\sqrt{(L+u_{n+1}-u_n)^2+(v_{n+1}-v_n)^2}\right)\right] ,$$

(A1)
where W is the potential energy of the stiff springs connecting the large masses and V is the potential energy of the weaker springs.

The associated equations of motion are

$$\begin{aligned} M\ddot{u}_n &= W'(L+u_{n+1}-u_n) - W'(L+u_n-u_{n-1}) \\ &\quad + S(D_{n+1})(L+u_{n+1}-u_n) - S(D_n)(L+u_n-u_{n-1}), \\ m\ddot{v}_n &= S(D_{n+1})(v_{n+1}-v_n) - S(D_n)(v_n-v_{n-1}) , \end{aligned}$$

(A2)
where we have set $S(D) := V'(D)/D$ and $D_n := \sqrt{(L+u_n-u_{n-1})^2+(v_n-v_{n-1})^2}$.

It is clear that one may take the undisturbed state of the internal modes $v_{n+1} = v_n$ for all n to hold exactly, in which case only the backbone motion is relevant. We will be interested in small amplitude, linear motions of the v_n, and how they are affected by the motion of the backbone.

The disparity between the masses M and m, and that of the strengths of the associated springs, is introduced by letting μ be a small parameter and assuming $m = \mu M$ and $V = \mu U$ (and correspondingly, $S = \mu Z$). We make the small-amplitude long-wave ansatz

$$u_n(t) = hu(X = hn, T = ht) \qquad \text{and} \qquad v_n(t) = hv(X = hn, T = ht)\,,$$

(A3)
where h is a small lattice-spacing parameter. Expanding the functions W and Z in Taylor series about the equilibrium position, the equations of motion become

$$\begin{aligned} Mh^3\partial_T^2 u = {}& h^3 W''(L)\partial_X^2 u + \frac{h^5}{12}W''(L)\partial_X^4 u \\ & + h^5 W'''(L)\partial_X u \cdot \partial_X^2 u + O(h^6) + O(h^3\mu)\,, \\ Mh^3\partial_T^2 v = {}& h^3 Z(L)\partial_X^2 v + \frac{h^5}{12}Z(L)\partial_X^4 v \\ & + h^5 Z'(L)\partial_X^2 u \cdot \partial_X v + h^5 Z'(L)\partial_X u \cdot \partial_X^2 v + O(h^6). \end{aligned}$$

(A4)
Trapping of v-waves by u-waves becomes possible if the wave speeds are equal. Therefore, we assume that $W''(L) = Z(L) = Mc^2$. Changing variables to $\chi = X - cT$ and $\tau = h^2 T$ yields

$$\begin{aligned} Mh^2\partial_\tau^2 u - 2Mc\partial_\chi\partial_\tau u &= \frac{Mc^2}{12}\partial_\chi^4 u + W'''(L)\partial_\chi u \cdot \partial_\chi^2 u + O(h) + O(\mu/h^2)\,, \\ Mh^2\partial_\tau^2 v - 2Mc\partial_\chi\partial_\tau v &= \frac{Mc^2}{12}\partial_\chi^4 v + Z'(L)\partial_\chi^2 u \cdot \partial_\chi v + Z'(L)\partial_\chi u \cdot \partial_\chi^2 v + O(h)\,. \end{aligned}$$

(A5)
As $h \downarrow 0$ with $\mu \ll h^2$, we find the coupled system

$$\begin{aligned} \partial_\tau A &+ \partial_\chi \left[\frac{1}{2}A^2 + \frac{c}{24}\partial_\chi^2 A\right] = 0\,, \\ \partial_\tau B &+ \partial_\chi \left[\frac{Z'(L)}{W'''(L)}AB + \frac{c}{24}\partial_\chi^2 B\right] = 0 \end{aligned}$$

(A6)
as a formal limit, where $A = W'''(L)\partial_\chi u/(2Mc)$ and $B = \partial_\chi v$. After a simple rescaling of χ and τ, this becomes (1) and (2) with $\kappa = Z'(L)/W'''(L)$. As described in section 6, the influence of solitons on linear waves can be qualitatively different for different values of the coupling constant κ with important bifurcations occurring at the values $\kappa = \kappa_n^{\mathrm{bif}} = (n+1)(n+2)/12$ for $n = 1, 2, 3, \ldots$. As we have seen, this coupled system can be solved exactly in (at least) two cases: $\kappa = 1$ and $\kappa = 1/2$. The former case is just the linearized KdV; see Sachs [S83]. The latter case is the one that is solved in the main text of this paper.

Consider this example with the potential of the strong and weak springs given respectively by

(A7)
$$W(D) := \frac{1}{2}\kappa_w D^2 + \frac{1}{24}\alpha D^4 \qquad \text{and} \qquad V(D) := \mu(\frac{1}{2}\kappa_v D^2 + \frac{1}{24}\beta D^4)\,.$$

Thus $\beta = 3\alpha\kappa$ and the condition that the wave speeds are equal is

(A8)
$$\frac{1}{2}\alpha L^2(\kappa - 1) = \kappa_w - \kappa_v\,.$$

The effect of each bifurcation point in κ is now clear. At $\kappa = 1/2$ the first harmonic of the v-waves begins to resonate with the u-waves. As κ increases through $\kappa = 1$ we pass through a transition from supercritical resonance to subcritical resonance. Similarly, at the odd bifurcation points, $\kappa = \kappa_{2m-1}^{\rm bif}$ for $m = 1, 2, 3, \ldots$, the mth harmonic v-wave begins to resonate with the u-waves. Then at the even bifurcation points, $\kappa = \kappa_{2m}^{\rm bif}$ for $m = 1, 2, 3, \ldots$, the nature of the resonance for this mode changes from supercritical to subcritical.

Coupled systems of equations like the pair (1) and (2) often arise as formal asymptotic reductions of mechanical models for complicated one-dimensional waves. Often these asymptotic models are integrable. For example, in an elastic rod, the interaction between axial twist waves and helical deformation waves gives rise to an integrable Manakov system of coupled nonlinear Schrödinger equations [LG99].

The coupled system (1) and (2) for $\kappa = 1/2$ is also intimately connected with an integrable multicomponent (an arbitrary number of components, all appearing symmetrically) coupled KdV equation [MC99]. Indeed, from one point of view it is this connection that yields the solvability of (1) and (2) for $\kappa = 1/2$ described in detail in this paper. The solution method presented here also can be used to give the complete solution of the coupled KdV system. That system in turn can be interpreted as a phenomenological model for the transport of the mass integral through an N-soliton solution of KdV [MC99].

Finally, we would like to point out that there are also some applications in which linear equations of the form (2) occur with $\kappa A(x, t)$ being a given function. In this case, $c(x, t) = \kappa A(x, t)$ represents a given spatiotemporal modulation of the speed of linear dispersive waves, say, due to propagation in an inhomogeneous medium. Such problems arise in the modeling of the propagation of weak internal waves in a channel of varying width [CG99]. For such applications, we may view the solvability of the coupled system (1) and (2) for $\kappa = 1$ and $\kappa = 1/2$ as a kind of (big) catalog of special cases of the function $c(x, t)$ for which the linear equation (2) is solvable in its own right. For other values of $\kappa > 1/2$ the linear wave system is unstable, while for all values of $\kappa < 1/2$ it is stable.

REFERENCES

[CG99] S. R. Clarke and R. H. J. Grimshaw, *Weakly nonlinear internal wave fronts trapped in contractions*, J. Fluid Mech., 415 (2000), pp. 323–345.

[CHQZ88] C. Canuto, M. Hussaini, A. Quateroni, and T. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, Berlin, 1988.

[GGKM67] C. Gardner, J. Greene, M. Kruskal, and R. Miura, *Method for solving the Korteweg-de Vries equation*, Phys. Rev. Lett., 19 (1967), pp. 1095–1097.

[GGKM74] C. Gardner, J. Greene, M. Kruskal, and R. Miura, *Korteweg-de Vries equation and generalizations*, VI. *Methods for exact solution*, Comm. Pure Appl. Math., 27 (1974), pp. 97–133.

[HS81] R. Hirota and J. Satsuma, *Soliton solutions of a coupled Korteweg-de Vries equation*, Phys. Lett., 85A (1981), pp. 407–408.

[KM56] I. Kay and H. E. Moses, *Reflectionless transmission through dielectrics and scattering potentials*, J. Appl. Phys., 27 (1956), pp.1503–1508.

[LG99] J. Lega and A. Goriely, *Pulses, fronts and oscillations of an elastic rod*, Phys. D, 132 (1999), pp. 373–391.

[MA98] P. D. Miller and N. N. Akhmediev, *Modal expansions and completeness relations for some time-dependent Schrödinger equations*, Phys. D, 123 (1998), pp. 513–524.

[MC99] P. D. Miller and P. L. Christiansen, *A coupled Korteweg-de Vries system and mass exchanges among solitons*, Phys. Scripta, 61 (2000), pp. 518–525.

[S83] R. L. Sachs, *Completeness of derivatives of squared Schrödinger eigenfunctions and explicit solutions of the linearized KdV equation*, SIAM J. Math. Anal., 14 (1983), pp. 674–683.

SIAM J. Math. Anal.
Vol. 33, No. 2, pp. 286–314

EXTENDING GEOMETRIC SINGULAR PERTURBATION THEORY TO NONHYPERBOLIC POINTS—FOLD AND CANARD POINTS IN TWO DIMENSIONS[*]

M. KRUPA[†][‡] AND P. SZMOLYAN[†]

Abstract. The geometric approach to singular perturbation problems is based on powerful methods from dynamical systems theory. These techniques have been very successful in the case of normally hyperbolic critical manifolds. However, at points where normal hyperbolicity fails, the well-developed geometric theory does not apply. We present a method based on blow-up techniques, which leads to a rigorous geometric analysis of these problems. A detailed analysis of the extension of slow manifolds past fold points and canard points in planar systems is given. The efficient use of various charts is emphasized.

AMS subject classifications. 34E15, 34E05, 34E20, 34C20, 34C26, 34C30, 34C40, 37C10

Key words. singular perturbations, slow manifolds, nonhyperbolicity, blow-up, folds, canards

PII. S0036141099360919

1. Introduction.

We consider singularly perturbed ordinary differential equations (ODEs) in the standard form

$$\begin{aligned} \varepsilon\dot{x} &= f(x,y,\varepsilon), \\ \dot{y} &= g(x,y,\varepsilon), \end{aligned} \qquad x \in \mathbb{R}^n, \quad y \in \mathbb{R}^m, \quad 0 < \varepsilon \ll 1, \tag{1.1}$$

where f, g are C^k-functions with $k \geq 3$. Properties of solutions of (1.1) can be studied using geometric methods from dynamical systems theory. This approach, known as *geometric singular perturbation theory*, has been very successful in many contexts, yet has encountered difficulties in certain situations. In this article we show how some of the limitations of geometric singular perturbation theory can be removed.

Before describing our results we present a brief survey of the existing theory. Let τ denote the independent variable in (1.1). The variable τ is referred to as the *slow* time scale. By switching to the *fast* time scale $t := \tau/\varepsilon$ one obtains the equivalent system

$$\begin{aligned} x' &= f(x,y,\varepsilon), \\ y' &= \varepsilon g(x,y,\varepsilon). \end{aligned} \tag{1.2}$$

One tries to analyze the dynamics of (1.1) by suitably combining the dynamics of the *reduced problem*

$$\begin{aligned} 0 &= f(x,y,0), \\ \dot{y} &= g(x,y,0) \end{aligned} \tag{1.3}$$

[*] Received by the editors September 1, 1999; accepted for publication (in revised form) December 16, 2000; published electronically June 8, 2001.

http://www.siam.org/journals/sima/33-2/36091.html

[†] Institut für Angewandte und Numerische Mathematik, Technische Universität Wien, Wiedner Hauptstraße 6-10, A-1040 Wien, Austria (peter.szmolyan@tuwien.ac.at). This research was supported by the Austrian Science Foundation under grant Y 42-MAT.

[‡] Current address: Department of Mathematical Sciences, New Mexico State University, Las Cruces, NM 88003-8001 (mkrupa@nmsu.edu).

and the dynamics of the *layer problem*

$$\begin{aligned} x' &= f(x,y,0), \\ y' &= 0, \end{aligned} \tag{1.4}$$

which are the limiting problems for $\varepsilon = 0$ on the slow and the fast time scales, respectively.

The foundation of geometric singular perturbation theory was laid by Fenichel [8]. The basic reasoning is as follows. The reduced problem (1.3) is a dynamical system on the set $S := \{(x,y) \in \mathbb{R}^{n+m} : f(x,y,0) = 0\}$. In the following we refer to S as the *critical manifold*. A normally hyperbolic invariant manifold of equilibria $S_0 \subset S$ of the layer problem (1.4) persists as a locally invariant slow manifold S_ε of (1.1) for ε sufficiently small. The restriction of (1.1) to S_ε is a small smooth perturbation of the reduced problem (1.3). Moreover, there exist a stable and an unstable invariant foliation with base S_ε with the dynamics along each foliation being a small perturbation of the suitable restriction of the dynamics of (1.4). For an excellent introduction to geometric singular perturbation theory and an overview of applications, we refer the reader to the survey by Jones [11].

However, despite many efforts, points on the critical manifold S where normal hyperbolicity breaks down remained a major obstacle to the geometric theory. This was a definite shortcoming in view of the abundance of nonhyperbolic points in applications.

One cause for the breakdown of normal hyperbolicity of a critical manifold S are bifurcation points due to a zero eigenvalue of the Jacobian $\frac{\partial f}{\partial x}$. The most common case are folded critical manifolds. A well-known phenomenon in this context are relaxation oscillations, i.e., solutions slowly moving towards a fold point, jumping from the fold point to another stable branch of S, following the slow dynamics again until another fold point is reached, jumping again, etc., thus, possibly forming periodic solutions [9], [18], and [20].

Another delicate phenomenon occuring at folds are *canard solutions* which were discovered and first analyzed by Benoit, Callot, Diener, and Diener [3]; see also [2]. A canard solution is a solution of a singularly perturbed system which is contained in the intersection of an attracting slow manifold and a repelling slow manifold. The existence of a canard solution can lead to *canard explosion*, i.e., a transition from a small limit cycle to a relaxation oscillation through a sequence of *canard cycles* [3], [6], [7]. For planar vector fields canards are nongeneric and occur persistently in one-parameter families; yet in dimensions larger than two they can occur in generic situations [2], [19], [23].

In this article we show how geometric singular perturbation theory can be extended to *fold points* and *canard points* in planar systems, i.e., we restrict our attention to the case $n = m = 1$. A fold point corresponds to the situation when the critical manifold has a generic fold. Depending on the stability properties of the critical manifold and on the direction of the reduced flow, a number of cases are possible. We analyze the so-called *jump point*, for which the reduced flow is directed towards the fold. This is the situation which is relevant for relaxation oscillations. We show how the slow manifolds (existing by the normally hyperbolic theory) extend in the neighborhood of the singularity. The treatment of the fold point is a refinement of the analysis in our earlier work [13]. A canard point is a fold point with an additional degeneracy leading to a possibility of a canard solution. Again we analyze how slow manifolds extend and show that a canard solution occurs along a codimension one curve in the parameter plane. In a complementary work [15] we carry out a similar analysis for singularities of pitchfork and transcritical type.

Our approach relies on the blow-up method, which is a way of partially desingularizing the vector field in the neighborhood of a singular point. After a blow-up transformation, standard methods from dynamical systems theory can be applied. Our proof of the existence of a canard solution is based on a variant of the Melnikov method. The blow-up method was first applied to a singular perturbation problem in the pioneering work of Dumortier and Roussarie [6], who analyzed the existence of *canard cycles* in the van der Pol equation. One of the purposes of this work is to build a bridge between the methods of [6] and geometric singular perturbation theory; in particular, we use the blow-up method to answer the question of extending slow manifolds near nonhyperbolic singularities.

The sequel of this article [14] is devoted to relaxation oscillations and canard explosion. The methods and results of this paper are of central importance in our analysis of these phenomena. Since the analysis of canard cycles is a more delicate problem, we find it advantageous to treat these issues separately. In particular, we prefer to use blow-up to obtain local results which in a second step can be used to study global phenomena. We feel that this point of view is also useful in the analysis of related problems. In this sense, our work is intended as a complement to the more global approach in the work of Dumortier [5] and Dumortier and Roussarie [6].

In this article and in [14] we restrict our attention to the planar case. However, the analysis carries over to higher-dimensional problems with one-dimensional critical manifolds containing fold points. By means of a center-manifold reduction, all normally hyperbolic directions can be eliminated and one recovers the planar problems considered here. A well-known problem where this is relevant is the traveling wave problem for the FitzHugh–Nagumo equation [11]. For a similar approach to problems with higher-dimensional critical manifolds, we refer the reader to [17] and [23].

The article is organized as follows. Section 2 contains the description and the analysis of a generic fold. Here we give a detailed expository presentation of the blow-up method. In section 3 we analyze a canard point.

2. Generic fold.

2.1. Assumptions and results. Consider the singularly perturbed ODE (1.2), where $(x, y) \in \mathbb{R}^2$ and ε is a small real parameter. Suppose that (x_0, y_0) is such that

$$f(x_0, y_0, 0) = 0, \qquad \frac{\partial f}{\partial x}(x_0, y_0, 0) = 0. \tag{2.1}$$

Our goal is to obtain a characterization of the dynamics in a neighborhood of (x_0, y_0) for sufficiently small values of ε. We make the following nondegeneracy assumptions:

$$\frac{\partial^2 f}{\partial x^2}(x_0, y_0, 0) \neq 0, \quad \frac{\partial f}{\partial y}(x_0, y_0, 0) \neq 0, \quad g(x_0, y_0, 0) \neq 0. \tag{2.2}$$

We assume, without loss of generality, that

$$(x_0, y_0) = (0, 0), \quad \frac{\partial^2 f}{\partial x^2}(0, 0, 0) > 0, \quad \frac{\partial f}{\partial y}(x_0, y_0, 0) < 0$$

hold.

As before let $S = \{(x, y) : f(x, y, 0) = 0\}$ be the critical manifold. The nondegeneracy assumptions imply that there exists a neighborhood U of the origin such that

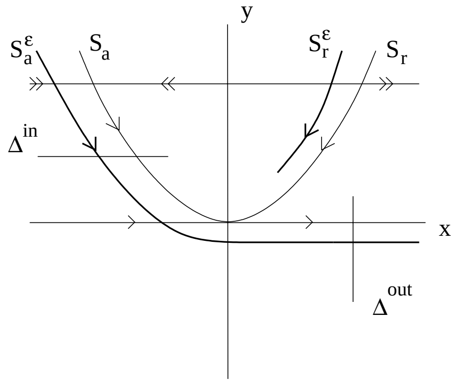


FIG. 2.1. *Critical manifold, slow manifolds, and sections for the fold point.*

$(0,0)$ is the only point in $U \cap S$, where $\frac{\partial f}{\partial x}$ vanishes and that $S \cap U$ is approximately a parabola. Let S_a (resp., S_r) denote its left (resp., right) branch, so that $S = S_a \cup S_r$ (see Figure 2.1). The assumption $\frac{\partial^2 f}{\partial x^2}(0,0,0) > 0$ implies that for $y > 0$ the branch S_a is attracting and the branch S_r is repelling for the layer problem, which also explains the notation. The origin is nonhyperbolic, weakly attracting from the left and weakly repelling to the right (see Figure 2.1).

To determine the reduced dynamics we solve the equation $f(x, y, 0) = 0$ for y as a function of x, i.e., $y = \varphi(x)$. The reduced dynamics is then determined by

$$\varphi'(x)\dot{x} = g(x, \varphi(x), 0), \tag{2.3}$$

which is singular at $x = 0$. Our assumptions on f and g imply that the direction of the reduced flow is determined by the sign of $g(0,0,0)$. We assume $g(0,0,0) < 0$. This implies that the reduced flow on S_a and S_r is directed towards the fold point; see Figure 2.1. Actually, orbits on S_a and S_r reach the fold point in finite time due to the singularity at the fold point. The only possibility to continue from there in the singular limit is along the (weakly) unstable fiber of the layer problem along the positive x-axis. Thus, the curve $S_a \cup \{(x, 0), x > 0\}$ is expected to be a zeroth order approximation. This is the situation relevant to relaxation oscillations; we refer to this case as *jump point*. The case $g(0,0,0) > 0$ can be analyzed similarly.

It follows from the standard theory [8] that outside an arbitrarily small neighborhood V of $(0,0)$, the manifolds S_a and S_r perturb smoothly to locally invariant manifolds $S_{a,\varepsilon}$ and $S_{r,\varepsilon}$ for sufficiently small $\varepsilon \neq 0$. We would like to point out that $S_{a,\varepsilon}$ and $S_{r,\varepsilon}$ are actually very simple. They consist of single solutions. Note that the slow manifolds are obtained as sections $\varepsilon =$ const. of two-dimensional, locally invariant, center-like manifolds M_a (resp., M_r) of the extended system

$$\begin{aligned} x' &= f(x, y, \varepsilon), \\ y' &= \varepsilon g(x, y, \varepsilon), \\ \varepsilon' &= 0 \end{aligned} \tag{2.4}$$

in the extended phase space $\mathbb{R}^3$. For this extended system $S \times \{0\}$ is a manifold of equilibria. Outside of a neighborhood of the fold point $(0,0,0)$ the linearization of system (2.4) at points $S_a \times \{0\}$ has a double zero eigenvalue and one uniformly hyperbolic (stable) eigenvalue. This allows us to conclude the existence of the attracting center-like manifold M_a; the manifold M_r is obtained in a similar way. At the fold point $(0,0,0)$ the linearization has a triple eigenvalue zero and the construction of the slow manifolds breaks down. We focus our attention on S_a and investigate how $S_{a,\varepsilon}$ as well as nearby solutions behave as they pass near the fold point. We expect that close to the fold point a transition from slow motion along $S_{a,\varepsilon}$ to a fast motion almost parallel to the unstable fibers occurs. A similar analysis could be carried out for $S_{r,\varepsilon}$.

Remark 2.1. It is known that the slow manifolds M_a and M_r and hence their sections $S_{a,\varepsilon}$ and $S_{r,\varepsilon}$ are not unique and are determined only up to $O(e^{-c/\varepsilon})$, where c is some positive constant. We make an arbitrary choice of M_a and M_r and indicate at the end that our results are independent of this choice.

We now view the previously introduced neighborhood U as a neighborhood of $(0,0,0)$ in $\mathbb{R}^3$. We pick U sufficiently small, so that $g(x,y,\varepsilon) \neq 0$ for $(x,y,\varepsilon) \in U$. Before stating the main results we rewrite system (1.2) (resp., (2.4)) in a canonical form. By rescaling x, y, ε, and t we obtain

$$\begin{aligned} x' &= -y + x^2 + h(x,y,\varepsilon), \\ y' &= \varepsilon g(x,y,\varepsilon), \\ \varepsilon' &= 0 \end{aligned} \tag{2.5}$$

with $h(x,y,\varepsilon) = O(\varepsilon, xy, y^2, x^3)$, $g(x,y,\varepsilon) = -1 + O(x,y,\varepsilon)$, where the new function g is related to the original one by the rescaling. This form of the equations will be used throughout the forthcoming analysis.

For small $\rho > 0$ and a suitable interval $J \subset \mathbb{R}$ let

$$\Delta^{in} = \{(x, \rho^2), x \in J\}$$

be a section in U transverse to S_a and let

$$\Delta^{out} = \{(\rho, y), y \in \mathbb{R}\}$$

be a section in U transverse to the fast fibers (see Figure 2.1). Note, that the same constant ρ is used throughout this paper.

Let $\pi : \Delta^{in} \to \Delta^{out}$ be the transition map for the flow of (1.2).

THEOREM 2.1. *Under the assumptions made in this section there exists $\varepsilon_0 > 0$ such that the following assertions hold for $\varepsilon \in (0, \varepsilon_0]$:*

1. *The manifold $S_{a,\varepsilon}$ passes through Δ^{out} at a point $(\rho, h(\varepsilon))$, where $h(\varepsilon) = O(\varepsilon^{2/3})$.*
2. *The transition map π is a contraction with contraction rate $O(e^{-c/\varepsilon})$, where c is a positive constant.*

In the context of matched asymptotic expansions assertion (1) of the theorem is well known; see, e.g., [18]. A blow-up based derivation of the asymptotic expansion of $h(\varepsilon)$ is given in [16]. Assertion (2) of the theorem explains why the nonuniqueness of the slow manifold M_a (resp., $S_{a,\varepsilon}$) does not affect our results. Two different choices of these manifolds are exponentially close at Δ^{in} and even more so at Δ^{out} due to the exponential contraction during the passage.

2.2. Blow-up. In this section we define and describe the blow-up transformation. The basic observation is that the fold point $(0,0,0)$ is a more degenerate equilibrium point of system (2.5) than the other points of the critical manifold S. The linearization of system (2.5) at the origin has a triple zero eigenvalue while the linearization at the other points of the critical manifold S has a double zero eigenvalue and one negative (resp., positive) eigenvalue for $x < 0$ (resp., $x > 0$).

The important insight in [6] is that blow-up techniques are the right tool to analyze nilpotent equilibria like the fold point, viewed as a degenerate equilibrium of the extended system (2.4). The blow-up method is essentially a clever coordinate transformation by which the degenerate equilibrium is "blown-up" to a two-sphere. In certain directions transverse to the sphere and even on the sphere, one gains enough hyperbolicity to allow a complete analysis by standard techniques. The technique is a generalization of the well known blow-up methods for degenerate equilibria of planar vector fields [5]. In the simplest situations this corresponds to blowing-up the degenerate equilibrium to the circle $r = 0$ by rewriting the vector field in polar coordinates $(r, \vartheta) \in \mathbb{R} \times S^1$. The analysis is often simplified substantially by using a quasi-homogeneous blow-up, i.e. by using different powers (weights) of r for different variables in the defining transformation.

The blow-up transformation for system (2.5) is

$$x = \bar{r}\bar{x}, \quad y = \bar{r}^2\bar{y}, \quad \varepsilon = \bar{r}^3\bar{\varepsilon} \tag{2.6}$$

with weights 1, 2, and 3. We define $B = S^2 \times [0, \rho]$, where the constant $\rho > 0$ is related to ε_0 by $\varepsilon_0 = \rho^3$. We consider the blow-up transformation as a mapping

$$\Phi : B \to \mathbb{R}^3 \tag{2.7}$$

with $(\bar{x}, \bar{y}, \bar{\varepsilon}) \in S^2$. We choose $\rho > 0$ sufficiently small such that system (2.4) is described by the canonical form (2.5) in the region $\Phi(B)$. We will be interested only in nonnegative values of $\bar{\varepsilon}$ and $\bar{r}$, but everything that follows makes sense for negative values as well, i.e., there are no technical problems at ∂B.

Let X denote the vector field corresponding to (2.5). Since X vanishes at the point $(0, 0, 0)$, there exists a vector field $\bar{X}$ on B such that $\Phi_* \bar{X} = X$, where Φ_* is induced by Φ. It remains to study the vector field $\bar{X}$ on the manifold B. Note that this suffices, since $\Phi(B)$ is a full neighborhood of the origin. In principle one could use spherical coordinates on S^2; however, this would lead to rather lengthy computations. It is natural and almost mandatory to use different charts for the manifold B to simplify the analysis. One reason for this is that—as we will see later—the dynamics in the individual charts is very different.

We will now introduce the charts used later in this paper. Loosely speaking, we will define a chart K_2, which describes a neighborhood of the upper half-sphere defined by $\bar{\varepsilon} > 0$, and charts K_1 and K_3 which describe neighborhoods of parts of the equator of S^2 which are needed in the analysis. In problems where a neighborhood of the whole equator needs to be analyzed, two further charts must be defined analogously. The subscripts in K_1, K_2, and K_3 denote the order in which the charts are used later.

The charts K_1, K_2, and K_3 are obtained by setting $\bar{y} = 1$, $\bar{\varepsilon} = 1$, and $\bar{x} = 1$, respectively, in the blow-up transformation (2.6). The blow-up transformation in the charts K_i, $i = 1, 2, 3$ is given by

$$x = r_1 x_1, \quad y = r_1^2, \quad \varepsilon = r_1^3 \varepsilon_1, \tag{2.8}$$

$$x = r_2 x_2, \quad y = r_2^2 y_2, \quad \varepsilon = r_2^3, \tag{2.9}$$

$$x = r_3, \quad y = r_3^2 y_3, \quad \varepsilon = r_3^3 \varepsilon_3 \tag{2.10}$$

with coordinates $(x_1, r_1, \varepsilon_1) \in \mathbb{R}^3$, $(x_2, y_2, r_2) \in \mathbb{R}^3$, and $(r_3, y_3, \varepsilon_3) \in \mathbb{R}^3$. The point $(0, 0, 0)$ is blown-up to the plane $r_i = 0$, $i = 1, 2, 3$. In our analysis we will need to change coordinates between these charts on their overlap domains. A simple computation gives the following lemma.

LEMMA 2.2. *Let κ_{12} denote the change of coordinates from K_1 to K_2. Then κ_{12} is given by*

$$x_2 = x_1 \varepsilon_1^{-1/3}, \quad y_2 = \varepsilon_1^{-2/3}, \quad r_2 = r_1 \varepsilon_1^{1/3} \quad \textit{for } \varepsilon_1 > 0, \tag{2.11}$$

and κ_{12}^{-1} is given by

$$x_1 = x_2 y_2^{-1/2}, \quad r_1 = r_2 y_2^{1/2}, \quad \varepsilon_1 = y_2^{-3/2} \quad \textit{for } y_2 > 0\ . \tag{2.12}$$

Let κ_{23} denote the change of coordinates from K_2 to K_3. Then κ_{23} is given by

$$r_3 = r_2 x_2, \quad y_3 = y_2 x_2^{-2}, \quad \varepsilon_3 = x_2^{-3} \quad \textit{for } x_2 > 0, \tag{2.13}$$

and κ_{23}^{-1} is given by

$$x_2 = \varepsilon_3^{-1/3}, \quad y_2 = y_3 \varepsilon_3^{-2/3}, \quad r_2 = r_3 \varepsilon_3^{1/3} \quad \textit{for } \varepsilon_3 > 0\ . \tag{2.14}$$

The above constructions make perfect sense if restricted to B. We introduce the following notation: $\bar{P}$ denotes an object in the blow-up which corresponds to an object P in the original problem. If $\bar{P}$ is described in one of the charts, then P_i denotes the object in chart K_i, $i = 1, 2, 3$. This notation is used only when necessary, mostly to denote various invariant manifolds.

Remark 2.2. In the work of Dumortier and Roussarie the chart K_2 corresponding to a directional blow-up in the direction of ε is called *family rescaling*, and charts used near the equator are called *phase directional rescaling.*

2.3. Blow-up of (1.2) with $\varepsilon = 0$. It is instructive to recall how the usual blow-up method applies to the layer problem, i.e., system (2.5) with $\varepsilon = 0$. Setting $\bar{\varepsilon} = 0$ in (2.6) defines a (planar, polar) blow-up of the degenerate equilibrium at the origin. To see this, note that $B \cap \{\bar{\varepsilon} = 0\} = S^1 \times [0, \rho]$, where $S^1 = \{(\bar{x}, \bar{y}, 0) \in S^2\}$. Due to the equation $\varepsilon' = 0$, the set $S^1 \times [0, \rho]$ is invariant for $\bar{X}$, which, restricted to $S^1 \times [0, \rho]$, is the blow-up of (1.2) with $\varepsilon = 0$.

Let $\bar{X}_0 = \bar{X}|_{S^1 \times [0,\rho]}$. Figure 2.2 shows the phase portrait of $\bar{X}_0$. We briefly describe this phase portrait, referring the reader to the sections on charts K_1 and K_3 for technical details. On the invariant circle S^1, there are four equilibria: p_a, p_r, q_{in}, q_{out}. These equilibria are hyperbolic for the flow on S^1, the points p_a and q_{out} are attracting, and p_r and q_{in} are repelling. The points p_a and p_r are end points of the blown-up critical manifolds $\bar{S}_a$ and $\bar{S}_r$, which are lines of equilibria for $\bar{X}_0$. Hence the radial direction is nonhyperbolic at p_a and p_r. The points q_{in} and q_{out} are the intersection points of S^1 with the blow-up of the critical fiber. These points are hyperbolic in the radial direction.

2.4. Dynamics in chart K_2. The dynamics of the blown-up vector field $\bar{X}$ in a neighborhood of the upper half-sphere is studied in chart K_2. The transformation (2.9) is just a rescaling of (x, y), since $r_2 = \varepsilon^{1/3}$. By inserting (2.9) into system (2.5) we obtain the vector field $\bar{X}$ in chart K_2. Since $r_2' = 0$, this blown-up system is still a

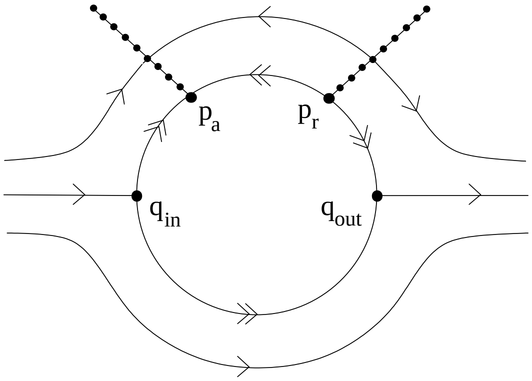


FIG. 2.2. *Phase portrait of the blown-up vector field for $\bar{\varepsilon} = 0$.*

family of planar vector fields with parameter r_2. We now desingularize the equations by rescaling time $t_2 := r_2 t$, so that the factor r_2 disappears. This desingularization is necessary to obtain a nontrivial flow on the blown-up locus $r_2 = 0$. We obtain

$$\begin{aligned} x_2' &= x_2^2 - y_2 + O(r_2), \\ y_2' &= -1 + O(r_2), \\ r_2' &= 0, \end{aligned} \tag{2.15}$$

where $'$ denotes differentiation with respect to t_2.

Remark 2.3. The rescaled form (2.15) of the original problem plays a crucial role in all approaches to the fold point by means of asymptotic expansions; e.g. [12], [18], and [20]. In these investigations solutions of (2.15) are used as inner solutions connecting (*matching*) solutions obtained as perturbations of the reduced problem to solutions obtained as solutions of the layer problem.

We first consider the case $r_2 = 0$, which gives

$$\begin{aligned} x_2' &= x_2^2 - y_2, \\ y_2' &= -1. \end{aligned} \tag{2.16}$$

This is a Riccati equation whose solutions can be expressed in terms of special functions. The relevant results can be found in [18, pp. 68–72]. Here we restate the results needed in our analysis. For the sake of readability we omit the subscript 2 of the variables.

PROPOSITION 2.3 (see [18]). *The Riccati equation* (2.16) *has the following properties:*

1. *Every orbit has a horizontal asymptote $y = y_r$, where y_r depends on the orbit such that $x \to \infty$ as y approaches y_r from above.*
2. *There exists a unique orbit γ_2 which can be parametrized as $(x, s(x))$, $x \in \mathbb{R}$ and is asymptotic to the left branch of the parabola $x^2 - y = 0$ for $x \to -\infty$. The orbit γ_2 has a horizontal asymptote $y = -\Omega_0 < 0$ such that $x \to \infty$ as y approaches $-\Omega_0$ from above.*
3. *The function $s(x)$ has the asymptotic expansions*

$$s(x) = x^2 + \frac{1}{2x} + O\left(\frac{1}{x^4}\right), \quad x \to -\infty,$$

$$s(x) = -\Omega_0 + \frac{1}{x} + O\left(\frac{1}{x^3}\right), \quad x \to \infty.$$

4. *All orbits to the right of γ_2 are backward asymptotic to the right branch of the parabola $x^2 - y = 0$.*
5. *All orbits to the left of γ_2 have a horizontal asymptote $y = y_l > y_r$, where y_l depends on the orbit, such that $x \to -\infty$ as y approaches y_l from below.*

Remark 2.4. The constant Ω_0 is the smallest positive zero of

$$J_{-1/3}(2z^{3/2}/3) + J_{1/3}(2z^{3/2}/3),$$

where $J_{-1/3}$ (resp., $J_{1/3}$) are Bessel functions of the first kind [18].

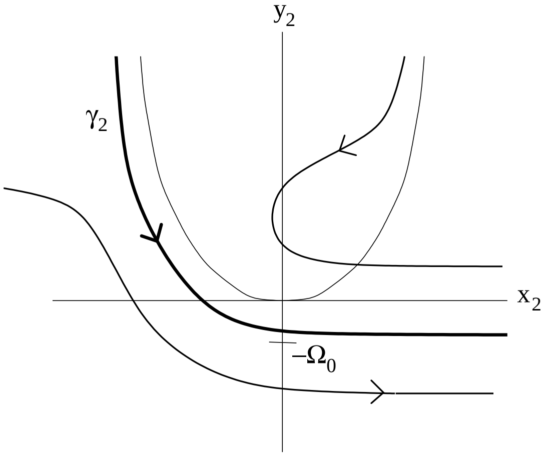


FIG. 2.3. *Solutions of the Riccati equation* (2.16).

The assertions of Proposition 2.3 are illustrated in Figure 2.3. We will see that the orbit $\bar{\gamma}$ corresponding to the special solution γ_2 is backward asymptotic to the equilibrium p_a on the equator of S^2. The importance of the orbit $\bar{\gamma}$ is that it "leads" the incoming attracting slow manifold across the upper half of the sphere S^2 to the point q_{out} from where take-off in the direction of the fast flow occurs.

We need to describe the transition map for (2.15) within a bounded domain D_2. Within such a domain we can deduce properties of the flow of (2.15) from Proposition 2.3 by using regular perturbation theory. A detailed study of the effect of the $O(r_2)$ perturbations outside D_2, i.e., close to infinity, will be carried out in the charts K_1 and K_3. For $\delta > 0$ we define the following sections:

$$\Sigma_2^{in} = \{(x_2, y_2, r_2) \,:\, y_2 = \delta^{-2/3}\}, \quad \Sigma_2^{out} = \{(x_2, y_2, r_2) \,:\, x_2 = \delta^{-1/3}\}.$$

Let Π_2 be the transition map of the flow (2.15) from Σ_2^{in} to Σ_2^{out}. Let $q_0 = \gamma_2 \cap \Sigma_2^{in}$.

PROPOSITION 2.4. *The transition map Π_2 has the following properties:*

1.

$$\Pi_2(q_0) = (\delta^{-1/3}, -\Omega_0 + \delta^{1/3} + O(\delta), 0).$$

2. *A neighborhood of q_0 is mapped diffeomorphically onto a neighborhood of $\Pi_2(q_0)$.*

Proof. The proof follows directly from Proposition 2.3 and regular perturbation theory. □

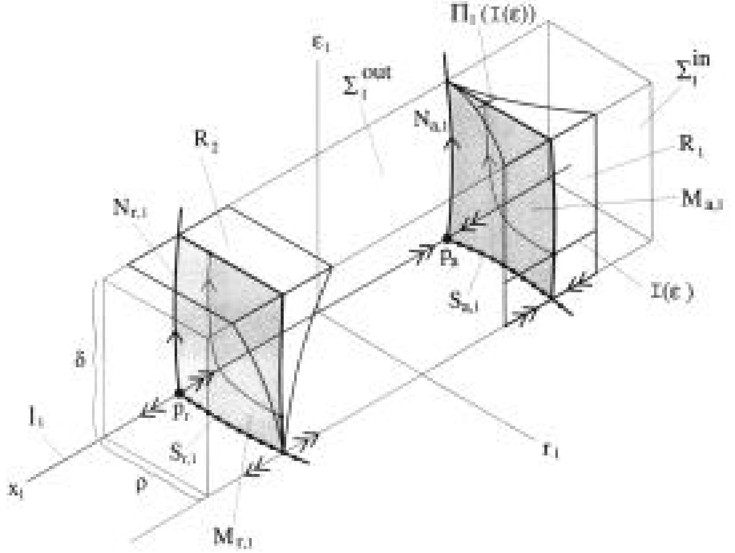


FIG. 2.4. *Geometry and dynamics in chart K_1.*

2.5. Dynamics in chart K_1. Chart K_1 is used to analyze the dynamics of the blown-up vector field $\bar{X}$ in a neighborhood of the equator containing the equilibria p_a and p_r. By inserting (2.8) into system (2.5), we obtain the vector field $\bar{X}$ in chart K_1. We desingularize the blown-up vector field $\bar{X}$ by dividing by r_1. This gives

$$\begin{aligned} x_1' &= -1 + x_1^2 + \frac{1}{2}\varepsilon_1 x_1 + O(r_1), \\ r_1' &= \frac{1}{2} r_1 \varepsilon_1(-1 + O(r_1)), \\ \varepsilon_1' &= \frac{3}{2}\varepsilon_1^2(1 + O(r_1)), \end{aligned} \tag{2.17}$$

where $'$ denotes differentiation with respect to a rescaled time variable t_1.

Remark 2.5. The equation for ε_1' is obtained from the equation $\varepsilon' = 0$, which implies the relation $3r_1^2 r_1' \varepsilon_1 + r_1^3 \varepsilon_1' = 0$. Hence $\varepsilon = r_1^3 \varepsilon_1$ is a constant of motion in chart K_1. Nevertheless, we will see that it is useful to treat the blown-up system as a three-dimensional problem. This seemingly artificial construction is actually crucial for the whole approach.

Remark 2.6. For $r_1 > 0$, system (2.17) has the same orbits as the blown-up vector field $\bar{X}$ with the corresponding solutions having a different time parametrization. Since we deal only with transition maps between sections, time parametrization of solutions has no significance to our analysis.

System (2.17) has two invariant subspaces, namely, the plane $r_1 = 0$ and the plane $\varepsilon_1 = 0$. Their intersection is the invariant line $l_1 := \{(x_1, 0, 0) \ : \ x_1 \in \mathbb{R}\}$; see Figure 2.4. The dynamics on l_1 is governed by $x_1' = -1 + x_1^2$. There are two equilibria $p_a = (-1, 0, 0)$ and $p_r = (1, 0, 0)$. For the flow on the line l_1 both points are hyperbolic, the relevant eigenvalue is -2 for p_a and 2 for p_r, i.e., p_a is attracting and p_r is repelling. The dynamics in the invariant plane $\varepsilon_1 = 0$ is governed by

$$\begin{aligned} x_1' &= -1 + x_1^2 + O(r_1), \\ r_1' &= 0. \end{aligned} \tag{2.18}$$

This system has a normally hyperbolic curve $S_{a,1}$ of equilibria emanating from p_a and a curve $S_{r,1}$ of equilibria emanating from p_r; see Figure 2.4. For r_1 small, this follows from the implicit function theorem. Actually, $S_{a,1}$ and $S_{r,1}$ are precisely the branches of the critical manifold S described in section 2.1; this also explains the notation. Along the curve $S_{a,1}$ the linearization of (2.18) has one zero eigenvalue, the other eigenvalue is negative and close to -2 for r_1 small. Along $S_{r,1}$ the situation is similar; however, the nonzero eigenvalue is positive and close to 2 for r_1 small.

Remark 2.7. Equation (2.18) is the directional (in the positive y-direction) blow-up of the fold point $(0, 0)$ of the layer problem, i.e., system (1.2) with $\varepsilon = 0$. The line $l_1 = 0$ corresponds to the fold point. We have gained normal hyperbolicity of the lines of equilibria $S_{a,1}$ (resp., $S_{r,1}$) at the points p_a (resp., p_r) due to the blow-up (compare Figure 2.2).

The dynamics in the invariant plane $r_1 = 0$ is governed by

$$\begin{aligned} x_1' &= -1 + x_1^2 + \frac{1}{2}\varepsilon_1 x_1, \\ \varepsilon_1' &= \frac{3}{2}\varepsilon_1^2\ . \end{aligned} \tag{2.19}$$

We recover the equilibria points p_a and p_r; however, there exists an additional zero eigenvalue due to the second equation. The corresponding eigenvector is $(-1, 4)$ at both equilibria. Hence, there exist one-dimensional center manifolds $N_{a,1}$ at p_a and $N_{r,1}$ at p_r along which ε_1 increases for $\varepsilon_1 > 0$. Note that the branch of the attracting center manifold $N_{a,1}$ at p_a in the half space $\varepsilon_1 > 0$ is unique, while the repelling center manifold $N_{r,1}$ at p_r in the half space $\varepsilon_1 > 0$ is not unique; see Figure 2.4. We collect the information we have obtained so far in the following lemma.

LEMMA 2.5. *The linearization of system* (2.17) *at* p_j, $j = a, r$ *has the following real eigenvalues:* $\lambda_1 = -2$ *at* p_a *and* $\lambda_1 = 2$ *at* p_r *with eigenvector* $(1, 0, 0)$ *corresponding to the flow on* l_1, $\lambda_2 = 0$ *with an eigenvector tangent to* $S_{j,1}$, *and* $\lambda_3 = 0$ *with an eigenvector* $(-1, 0, 4)$ *corresponding to the center direction in the invariant plane* $r_1 = 0$.

We restrict our attention to the set

$$D_1 := \{(x_1, r_1, \varepsilon_1) : \ x_1 \in \mathbb{R}, 0 \le r_1 \le \rho,\ 0 \le \varepsilon_1 \le \delta\},$$

where $\rho > 0$ is the constant defining the sections Δ^{in} and Δ^{out} in section 2.1 and $\delta > 0$ is the constant defining the sections Σ_2^{in} and Σ_2^{out} in section 2.4. Note that all objects defined later extend smoothly to negative values of r_1 and ε_1; i.e., there are no problems due to the boundaries $r_1 = 0$ and $\varepsilon_1 = 0$. We have the following result.

PROPOSITION 2.6. *For* ρ, δ *sufficiently small the following assertions hold for system* (2.17):

1. *There exists an attracting two-dimensional* C^k*-center manifold* $M_{a,1}$ *at* p_a *which contains the line of equilibria* $S_{a,1}$ *and the center manifold* $N_{a,1}$. *In* D_1 *the manifold* $M_{a,1}$ *is given as a graph* $x_1 = h_a(r_1, \varepsilon_1)$. *The branch of* $N_{a,1}$ *in* $r_1 = 0$, $\varepsilon_1 > 0$ *is unique.*

2. *There exists a repelling two-dimensional C^k-center manifold $M_{r,1}$ at p_r which contains the line of equilibria $S_{r,1}$ and the center manifold $N_{r,1}$. In D_1 the manifold $M_{r,1}$ is given as a graph $x_1 = h_r(r_1, \varepsilon_1)$. The branch of $N_{r,1}$ in $r_1 = 0$, $\varepsilon_1 > 0$ is not unique.*
3. *There exists a stable invariant foliation $\mathcal{F}^s$ with base $M_{a,1}$ and one-dimensional fibers. For any $c > -2$ there exists a choice of positive ρ and δ such that the contraction along $\mathcal{F}^s$ during a time interval $[0, T]$ is stronger than e^{cT}.*
4. *There exists an unstable invariant foliation $\mathcal{F}^u$ with base $M_{r,1}$ and one-dimensional fibers. For any $c < 2$ there exists a choice of positive ρ and δ such that the expansion along $\mathcal{F}^u$ during a time interval $[0, T]$ is stronger than e^{cT}.*
5. *The unique branch of $N_{a,1}$ in $r_1 = 0$, $\varepsilon_1 > 0$ is equal to $\gamma_1 := \kappa_{12}^{-1}(\gamma_2)$, wherever κ_{12}^{-1} is defined, i.e., along the part of γ_2 corresponding to $y_2 > 0$.*

Proof. Assertions (1)–(4) follow from Lemma 2.5 and center manifold theory; see, e.g., [4], [10]. Proposition 2.3 and the coordinate transformation (2.12) imply that $\kappa_{12}^{-1}(\gamma_2)$ has the expansion

$$\left(x_2 \left(x_2^2 + \frac{1}{2x_2} + O\left(\frac{1}{x_2^4}\right)\right)^{-1/2},\ 0,\ \left(x_2^2 + \frac{1}{2x_2} + O\left(\frac{1}{x_2^4}\right)\right)^{-3/2} \right)$$

as $x_2 \to -\infty$. Expanding these terms in powers of x_2 shows that $\kappa_{12}^{-1}(\gamma_2)$ converges to p_a tangent to the center-direction $(-1, 0, 4)$ as $x_2 \to -\infty$. This and the uniqueness of the branch of $N_{a,1}$ in $r_1 = 0$, $\varepsilon_1 > 0$ imply assertion (5). □

Remark 2.8. Clearly, the center manifold $M_{a,1}$ in chart K_1 corresponds to a locally invariant manifold $\bar{M}_a$ of the blown-up vector field $\bar{X}$. The importance of assertion (5) in the above proposition is that it allows us to track the manifold $\bar{M}_a$ as it moves across the sphere S^2 guided by the special orbit $\bar{\gamma}$ corresponding to the solution γ_2 of the Riccati equation.

We now define the following sections:

$$\Sigma_1^{in} := \{(x_1, r_1, \varepsilon_1) \in D_1 \ :\ r_1 = \rho\}, \quad \Sigma_1^{out} := \{(x_1, r_1, \varepsilon_1) \in D_1 \ :\ \varepsilon_1 = \delta\}.$$

Remark 2.9. Note that Σ_1^{in} maps under the blow-up transformation (2.8) to Δ^{in} and Σ_1^{out} maps under the coordinate transformation (2.11) to Σ_2^{in}. An important part of our description of the flow near the fold is the description of the transition map from Σ_1^{in} to Σ_1^{out} near the center manifolds $M_{a,1}$ and $M_{r,1}$. Since the neighborhood of $M_{a,1}$ corresponds to the neighborhood of the attracting branch of the slow manifold of (1.2), we are more interested in understanding the dynamics near $M_{a,1}$. Yet the analysis of the two cases is very similar, so we handle them simultaneously.

Let R_1 be the rectangle in Σ_1^{in} defined by $|1+x_1| \le \beta_1$, and let R_2 be the rectangle in Σ_1^{out} defined by $|1 - x_1| \le \beta_1$ for sufficiently small $\beta_1 > 0$. The constants ρ, δ, and β_1 can be chosen such that $M_{a,1} \cap \Sigma_1^{in} \subset R_1$ and $M_{r,1} \cap \Sigma_1^{out} \subset R_2$. For $0 \le \tilde{\varepsilon} \le \delta$ and $0 \le \tilde{r} \le \rho$, let $I_a(\tilde{\varepsilon})$ be the line $R_1 \cap \{\varepsilon_1 = \tilde{\varepsilon}\}$ and $I_r(\tilde{r})$ be the line $R_2 \cap \{r_1 = \tilde{r}\}$.

In the neighborhood of p_j, $j = a, r$, the flow of (2.17) carries Σ_1^{in} to Σ_1^{out}. Let $\Pi_1 : \Sigma_1^{in} \to \Sigma_1^{out}$ be the transition map defined by the flow of (2.17). The map Π_1 is well defined on R_1, at least for small enough values of ρ, δ, and β_1. The map Π_1 is defined in a wedge-shaped set in Σ_1^{in} around $M_{r,1}$ that shrinks to $S_{r,1}$ for $\varepsilon_1 \to 0$. The reason for this difference is that $M_{a,1}$ is attracting and $M_{r,1}$ is repelling. We have the following estimate of transition times.

LEMMA 2.7. *The transition time T of a solution of system* (2.17) *from a point $p = (x_1, \rho, \varepsilon_1) \in \Sigma_1^{in}$ to the point $\Pi_1(p) \in \Sigma_1^{out}$ satisfies*

$$T = \frac{2}{3}\left(\frac{1}{\varepsilon_1} - \frac{1}{\delta}\right)(1 + O(\rho)) . \tag{2.20}$$

Proof. The evolution of ε_1 determines the transition time of solutions from Σ_1^{in} to Σ_1^{out}. The relevant equation is

$$\varepsilon_1' = \frac{3}{2}\varepsilon_1^2(1 - O(r_1)) . \tag{2.21}$$

The result follows immediately by integrating (2.21). □

PROPOSITION 2.8. *For ρ, δ, and β_1 sufficiently small the transition map $\Pi_1 : \Sigma_1^{in} \to \Sigma_1^{out}$ defined by the flow of system* (2.17) *has the following properties:*

1. *$\Pi_1(R_1)$ is a wedge-like region in Σ_1^{out}. $\Pi_1^{-1}(R_2)$ is a wedge-like region in Σ_1^{in}.*
2. *More precisely, for fixed $c < 2$ there exists a constant K depending on the constants c, ρ, δ, and β_1 such that*
   (i) *for $\varepsilon_1 \in (0, \delta]$ the map $\Pi_1|I_a(\varepsilon_1)$ is a contraction with contraction rate bounded by $Ke^{-\frac{2c}{3}\left(\frac{1}{\varepsilon_1}-\frac{1}{\delta}\right)}$.*
   (ii) *for $r_1 \in (0, \rho]$ the map $\Pi_1^{-1}|I_r(r_1)$ is a contraction with contraction rate bounded by $Ke^{-\frac{2c}{3}\left(\frac{\rho^3}{r_1^3\delta}-\frac{1}{\delta}\right)}$.*

Proof. The assertions follow from Proposition 2.6 and Lemma 2.7. The estimate for the contraction rate of Π_1^{-1} in the second assertion uses the identity $\varepsilon_1\rho^3 = \delta r_1^3$ for $p = (x_{1,in}, \rho, \varepsilon_1)$ and $\Pi_1(p) = (x_{1,out}, r_1, \delta)$ to express the transition time in terms of r_1. □

All our results concerning the dynamics in chart K_1 are illustrated in Figure 2.4.

2.6. Dynamics in chart K_3. We use chart K_3 to analyze the dynamics of the blown-up vector field $\bar{X}$ in a neighborhood of the equator containing the point q_{out}. Applying transformation (2.10) to system (2.5) and desingularizing by dividing out the factor r_3, we obtain

$$\begin{aligned} r_3' &= r_3F(r_3, y_3, \varepsilon_3), \\ y_3' &= \varepsilon_3(-1 + O(r_3)) - 2y_3F(r_3, y_3, \varepsilon_3), \\ \varepsilon_3' &= -3\varepsilon_3F(r_3, y_3, \varepsilon_3), \end{aligned} \tag{2.22}$$

where $F(r_3, y_3, \varepsilon_3) := 1 - y_3 + O(r_3)$. The planes $\varepsilon_3 = 0$ and $r_3 = 0$ and the y_3-axis are invariant under the flow of (2.22).

LEMMA 2.9. *The point $q_{out} = (0, 0, 0)$ is a hyperbolic equilibrium of system* (2.22) *with eigenvalues: $\lambda_1 = 1$ with eigenvector $(1, 0, 0)$ corresponding to the flow in $\varepsilon_3 = 0$, $\lambda_2 = -2$ with eigenvector $(0, 1, 0)$ corresponding to the flow on the y_3-axis, and $\lambda_3 = -3$ with eigenvector $(0, 1, 1)$ corresponding to the flow in $r_3 = 0$.*

Proof. Computation. □

Now we transform the part of the special orbit γ_2 (introduced in Proposition 2.3) corresponding to $x_2 > 0$ to chart K_3; i.e., we define $\gamma_3 := \kappa_{23}(\gamma_2)$.

LEMMA 2.10. *The orbit γ_3 lies in the plane $r_3 = 0$, converges to q_{out} as $\varepsilon_3 \to 0$, and is tangent at q_{out} to the vector $(0, 1, 0)$.*

Proof. The coordinate transformation (2.13) and assertion (3) from Proposition 2.3 imply that the orbit γ_3 has the expansion $(0,\ -\Omega_0\varepsilon_3^{2/3} + \varepsilon_3 + O(\varepsilon_3^{5/3}),\ \varepsilon_3\ )$ as $\varepsilon_3 \to 0$. The lemma follows. □

Lemma 2.10 implies that parts of the manifold $\bar{M}_a$ corresponding to $\bar{r} > 0$ come close to the equilibrium q_{out}. Hence, we need a precise description of the dynamics of system (2.22) close to q_{out}. This is a somewhat delicate problem because of the *resonance* $\lambda_2 = \lambda_1 + \lambda_3$, which implies that there exists no smooth transformation of the nonlinear flow to the flow of the corresponding linearization. It turns out that, due to the simple form of the equations, it is quite easy to work out the lowest order approximation of the flow.

For the description of the flow in a neighborhood of q_{out} we define sections Σ_3^{in} and Σ_3^{out} as follows:

$$\Sigma_3^{in} = \{(r_3, y_3, \varepsilon_3) \;:\; r_3 \in [0, \rho],\, y_3 \in [-\beta_3, \beta_3],\, \varepsilon_3 = \delta\},$$
$$\Sigma_3^{out} = \{(r_3, y_3, \varepsilon_3) \;:\; r_3 = \rho,\, y_3 \in [-\beta_3, \beta_3],\, \varepsilon_3 \in [0, \delta]\},$$

where ρ and δ are the same constants as before, and $\beta_3 > 0$ is sufficiently small; see Figure 2.5.

Let Π_3 be the transition map from Σ_3^{in} to Σ_3^{out}. Our goal is to obtain a formula for the map Π_3. Before stating the relevant result we need to discuss the structure of (2.22) in more detail. We first divide (2.22) by the factor $F(r_3, y_3, \varepsilon_3)$, which is close to one near q_{out}, and obtain

$$\begin{aligned} r_3' &= r_3, \\ y_3' &= -2y_3 - \frac{\varepsilon_3}{1 - y_3} + r_3\varepsilon_3 G(r_3, y_3, \varepsilon_3), \\ \varepsilon_3' &= -3\varepsilon_3, \end{aligned} \tag{2.23}$$

where $G(r_3, y_3, \varepsilon_3)$ is a C^k-function. Consider (2.23) with $r_3 = 0$, namely,

$$\begin{aligned} y_3' &= -2y_3 - \frac{\varepsilon_3}{1 - y_3}, \\ \varepsilon_3' &= -3\varepsilon_3. \end{aligned} \tag{2.24}$$

By construction, system (2.24) is, up to rescaling of time, the Riccati equation (2.16) transformed to K_3. The corresponding linearization has eigenvalues $\lambda_2 = -2$ and $\lambda_3 = -3$; hence, (2.16) can be linearized by a near identity transformation of the form

$$y_3 = \psi(\tilde{y}_3, \varepsilon_3), \tag{2.25}$$

where the function ψ is C^k smooth and $\psi(\tilde{y}_3, \varepsilon_3) = \tilde{y}_3 + O(\tilde{y}_3\varepsilon_3)$; see [22]. The corresponding inverse transformation is denoted by $\tilde{y}_3 = \tilde{\psi}(y_3, \varepsilon_3) = y_3 + O(y_3\varepsilon_3)$. Under the transformation (2.25) system (2.23) becomes

$$r_3' = r_3, \tag{2.26a}$$
$$\tilde{y}_3' = -2\tilde{y}_3 - \varepsilon_3 + r_3\varepsilon_3 H(r_3, \tilde{y}_3, \varepsilon_3), \tag{2.26b}$$
$$\varepsilon_3' = -3\varepsilon_3, \tag{2.26c}$$

with a C^k-function H. We have the following result.

Proposition 2.11. *The transition map* Π_3 *for system* (2.22) *has the form*

$$\Pi_3(r_3, y_3, \delta) = \begin{pmatrix} \rho \\ \Pi_{32}(r_3, y_3, \delta) \\ \left(\frac{r_3}{\rho}\right)^3 \delta \end{pmatrix}$$

with $\Pi_{32}(r_3, y_3, \delta)$ *given by*

$$\Pi_{32}(r_3, y_3, \delta) = (\tilde{\psi}(y_3, \delta) - \delta)\left(\frac{r_3}{\rho}\right)^2 + O(r_3^3 \ln r_3).$$

Proof. In the following we suppress the subscript 3 in system (2.26). Fix $(r_i, \tilde{y}_i, \delta) \in \Sigma^{in}$ and $(\rho, \tilde{y}_o, \varepsilon_o) \in \Sigma^{out}$. Consider a solution $(r, \tilde{y}, \varepsilon)(t)$ of (2.26) and $T > 0$ such that $r(0) = r_i$, $r(T) = \rho$, $\tilde{y}(0) = \tilde{y}_i$, $\tilde{y}(T) = \tilde{y}_o$, $\varepsilon(0) = \delta$, $\varepsilon(T) = \varepsilon_o$. We will now compute $(T, \tilde{y}_o, \varepsilon_o)$ as a function of $(r_i, \tilde{y}_i)$. Equations (2.26a) and (2.26c) have explicit solutions $r = e^t r_i$, $\varepsilon = \delta e^{-3t}$. The requirement $r(T) = \rho$ produces an expression for T, namely,

$$T = \ln\left(\frac{\rho}{r_i}\right). \tag{2.27}$$

Let z be a new coordinate defined by $\tilde{y} = e^{-2t}(\tilde{y}_i - \delta + z) + \delta e^{-3t}$. We get the following equation for z:

$$z' = r_i H^z(z, r_i, \tilde{y}_i, t), \tag{2.28}$$

where $H^z(z, r_i, \tilde{y}_i, t) = \delta H(e^t r_i, e^{-2t}(\tilde{y}_i - \delta + z) + \delta e^{-3t}, \delta e^{-3t})$. The transition time T is still given by (2.27). Note that the function H^z is uniformly bounded on the relevant domain. Using (2.28) we obtain $z(T) = r_i O(T) = O(r_i \ln(\frac{\rho}{r_i}))$. It follows that

$$\tilde{y}(T) = (\tilde{y}_i - \delta)\left(\frac{r_i}{\rho}\right)^2 + O\left(\frac{r_i^3}{\rho^2} \ln\left(\frac{\rho}{r_i}\right)\right). \tag{2.29}$$

Hence

$$\begin{aligned}\Pi_{32}(r_3, y_3, \delta) &= \psi\left((\tilde{\psi}(y_3, \delta) - \delta)\left(\frac{r_3}{\rho}\right)^2 + O(r_3^3 \ln r_3), \left(\frac{r_3}{\rho}\right)^3 \delta\right)\\ &= (\tilde{\psi}(y_3, \delta) - \delta)\left(\frac{r_3}{\rho}\right)^2 + O(r_3^3 \ln r_3)\end{aligned}$$

and the result follows. □

Remark 2.10. The following observation will be used later in this paper to obtain the leading order asymptotics of the extended slow manifold $S_{a,\varepsilon}$. The y_3 coordinate of the point where the special orbit γ_3 intersects the section Σ_3^{in} is $y_3^* = \delta^{2/3} s(\delta^{-1/3})$ (see the proof of Lemma 2.10). By comparing the asymptotics of γ_3 and the exact solution of system (2.26) restricted to $r_3 = 0$, i.e., the Riccati equation written in the linearizing coordinates $(\tilde{y}_3, \varepsilon_3)$, it follows that $\tilde{\psi}(y_3^*, \delta) - \delta = -\Omega_0 \delta^{2/3}$.

2.7. Phase portrait on the upper part of S^2. The sphere S^2 is invariant under the desingularization of the blown-up vector field $\bar{X}$. The equator S^1 is invariant. On S^1 there are four equilibria p_a, p_r, q_{in}, q_{out}. These equilibria are hyperbolic for the flow on S^1, the points p_a and q_{out} are attracting, and p_r and q_{in} are repelling. All orbits in $S^{2,+}$ are forward asymptotic to q_{out}. The special orbit $\bar{\gamma}$ is backward asymptotic to p_a and, as it arrives at q_{out}, it is tangent to S^1. Besides $\bar{\gamma}$ there exist two families of trajectories: backward asymptotic to p_r or backward asymptotic to q_{in}. The corresponding phase portrait of S^2 is shown in Figure 2.6.

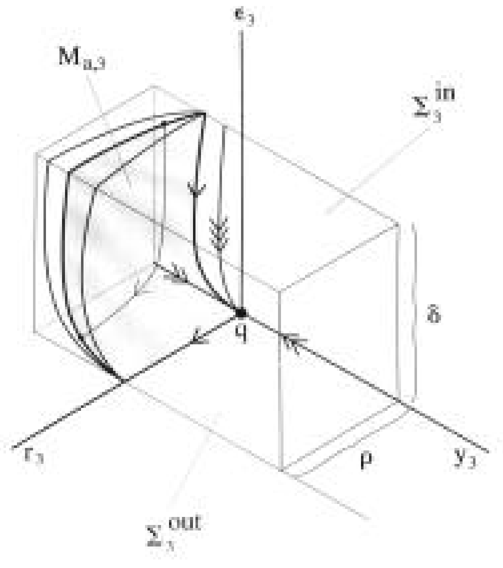


FIG. 2.5. *Geometry and dynamics of system* (2.22) *near the equilibrium* q_{out}.

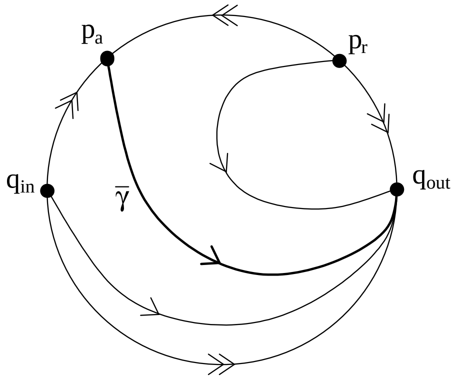


FIG. 2.6. *Blow-up of a fold (jump) point restricted to* S^2.

2.8. Proof of the main result. In this section we prove Theorem 2.1 by combining the results obtained in the individual charts. The idea of the proof is to analyze the evolution of the center manifold $M_{a,1}$ and the rectangle R_1 under the flow of the blown-up vector field $\bar{X}$. We denote the corresponding global invariant manifold by $\bar{M}_a$. The intersection of $\bar{M}_a$ with $S^{2,+}$ is the special orbit $\bar{\gamma}$ connecting the equilibria p_a and q_{out}. It follows that a trajectory starting in or nearby $\bar{M}_a$ will remain close to $\bar{\gamma}$ until it reaches the vicinity of q_{out}. There the trajectory follows the local dynamics near q_{out} and is repelled in the unstable direction of q_{out}.

Proof of Theorem 2.1. We define the map $\Pi : \Sigma_1^{in} \to \Sigma_3^{out}$ by

$$\Pi := \Pi_3 \circ \kappa_{23} \circ \Pi_2 \circ \kappa_{12} \circ \Pi_1.$$

The map Π is the transition map from Σ_1^{in} to Σ_3^{out} for the flow induced by the blown-up vector field $\bar{X}$ on B. We will analyze $\Pi(R_1 \cap M_{a,1})$ and then use the fact that, by construction, the transition map π is given by $\pi = \Phi \circ \Pi \circ \Phi^{-1}$ for $\varepsilon > 0$.

It follows from Proposition 2.8 that $\Pi_1(R_1 \cap M_{a,1}) \subset \Sigma_1^{out}$ is a smooth curve transverse to the set $\{r_1 = 0\}$. It follows that $\kappa_{12}(\Pi_1(R_1 \cap M_{a,1}))$ is a smooth curve transverse to the plane $\{r_2 = 0\}$. Proposition 2.4 implies that the image of this curve under Π_2 has the form $\{(\delta^{-1/3}, h_2^{out}(r_2), r_2) \ : \ r_2 \in [0, \rho\delta^{1/3}]\}$, where $h_2^{out} : [0, \rho\delta^{1/3}] \to \mathbb{R}$ is a smooth function. Under the transformation κ_{23}, this curve transforms to a smooth curve of the form $\{(r_3, h_3^{in}(r_3), \delta) : r_3 \in [0, \rho]\}$ with $(0, h_3^{in}(0), \delta) = \kappa_{23}(\gamma_2 \cap \Sigma_2^{out})$. Proposition 2.11 now implies that $\Pi(R_1 \cap M_{a,1})$ has the form $\{(\rho, h_3^{out}(\varepsilon_3), \varepsilon_3) : \varepsilon_3 \in [0, \delta]\}$, where $h_3^{out}(\varepsilon_3) = O(\varepsilon_3^{2/3})$. This proves assertion (1) of the theorem.

We now prove assertion (2). It follows from Proposition 2.8 that $\Pi_1(R_1)$ is a wedge-like region around $\Pi(R_1 \cap M_{a,1})$ of width $O(e^{-c/\varepsilon_1})$, where $c > 0$ is some constant. Since κ_{12}, Π_2 and κ_{23} are diffeomorphisms restricted to Σ_1^{out}, Σ_2^{in}, and Σ_2^{out}, respectively, it follows that $\kappa_{23} \circ \Pi_2 \circ \kappa_{12} \circ \Pi_1(R_1)$ is also a wedge-like region of width $O(e^{-c/\varepsilon_1})$ around $\kappa_{23} \circ \Pi_2 \circ \kappa_{12} \circ \Pi_1(R_1 \cap M_{a,1})$. Finally, we apply Proposition 2.11 to conclude that $\Pi(R_1)$ is a wedge-like region of width $O(e^{-c/\varepsilon_1})$ around $\Pi(R_1 \cap M_{a,1})$. The evolution of R_1 in the three charts is shown in Figure 2.7. Because $\varepsilon = \rho^3\varepsilon_1 = \rho^3\varepsilon_3$ is a constant of motion for the flow of $\bar{X}$, lines $\varepsilon_1 = \varepsilon/\rho^3$ in Σ_1^{in} are mapped to lines $\varepsilon_3 = \varepsilon/\rho^3$ in Σ_3^{out}. Restricted to such lines the map Π is a contraction with contraction rate $O(e^{-c/\varepsilon_1})$ for some $c > 0$. Assertion (2) follows by applying the appropriate blow-down transformations. □

The dynamics of the blown-up vector field $\bar{X}$, and in particular the center manifold $\bar{M}_a$, are shown in Figure 2.7.

Remark 2.11. Remark 2.10 implies that the function $h_3^{out}(\varepsilon_3)$ has the asymptotic expansion

$$h_3^{out}(\varepsilon_3) = -\Omega_0\varepsilon_3^{2/3} + o(\varepsilon_3^{2/3}).$$

The corresponding expansion for the function $h(\varepsilon)$ in Theorem 2.1 is

$$h(\varepsilon) = -\Omega_0\varepsilon^{2/3} + o(\varepsilon^{2/3}).$$

This result is well known; see, e.g., [18], where it is also shown that the next term in the expansion is $O(\varepsilon \ln \varepsilon)$. Our analysis, in particular the description of the map Π_3 in Proposition 2.11, shows that the occurrence of this term is due to the resonance $\lambda_2 = \lambda_1 + \lambda_3$ at the equilibrium q_{out}; see section 2.6.

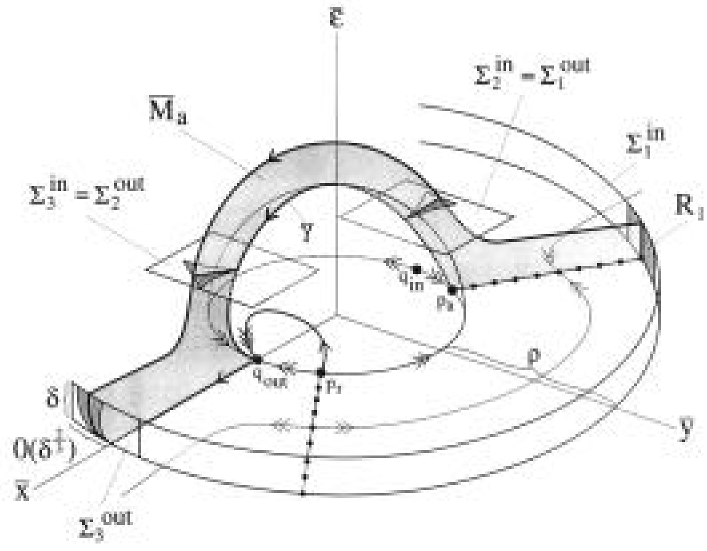


FIG. 2.7. *Geometry and dynamics of the blown-up vector field $\bar{X}$.*

3. Canard point.

3.1. Assumptions and results. In this section we consider a one parameter family of ODEs similar to (1.2), namely,

$$\begin{aligned} x' &= f(x, y, \lambda, \varepsilon), \\ y' &= \varepsilon g(x, y, \lambda, \varepsilon). \end{aligned} \tag{3.1}$$

We assume that $(x_0, y_0) = (0, 0)$ is a nondegenerate fold point of the critical manifold $f(x, y, \lambda, 0) = 0$ for $\lambda_0 = 0$. We assume further that $g(0, 0, 0, 0) = 0$. This gives the following set of defining conditions for the considered singularity:

$$f(0,0,0,0) = 0, \quad \frac{\partial f}{\partial x}(0,0,0,0) = 0, \quad g(0,0,0,0) = 0 \tag{3.2}$$

with the nondegeneracy assumptions

$$\frac{\partial^2 f}{\partial x^2}(0,0,0,0) \neq 0, \quad \frac{\partial f}{\partial y}(0,0,0,0) \neq 0. \tag{3.3}$$

This implies that the critical manifold has a nondegenerate fold point for λ in a suitable interval. Without loss of generality we assume that the fold point is $(0, 0)$ for all values of λ. This can always be achieved by a λ-dependent translation. The remaining nondegeneracy assumptions defining a canard point are

$$\frac{\partial g}{\partial x}(0,0,0,0) \neq 0, \quad \frac{\partial g}{\partial \lambda}(0,0,0,0) \neq 0\ . \tag{3.4}$$

These conditions insure that the nullcline $g(x, y, \lambda, 0) = 0$ is transverse to the critical manifold S and the intersection point of S and $g(x, y, \lambda, 0) = 0$ passes through the fold point $(0, 0)$ with nonzero speed as λ varies.

Let the critical manifold S, its left and right branches S_a and S_r, and the neighborhoods U and V be defined as in section 2. The manifolds $S_{a,\varepsilon}$ and $S_{r,\varepsilon}$ exist outside of V just as they did for a simple fold. Here we ask basically the same question, namely, How can S_a and S_r be extended? The situation is, however, quite different, since, for special choices of λ and ε, $S_{a,\varepsilon}$ extends to $S_{r,\varepsilon}$. This is caused by the special structure of the slow flow for $\lambda = 0$.

As in the case of the fold point, the reduced dynamics is governed by the equation

$$\dot{x} = \frac{g(x, \varphi(x), 0, 0)}{\varphi'(x)}, \tag{3.5}$$

where $\varphi(x)$ is obtained by solving the equation $f(x, y, 0, 0)$ for y as a function of x. It follows from the above assumptions that the right-hand side of (3.5) is a smooth function at the origin. Let $x_0(t)$ denote a maximal solution of (3.5) with the property that $x_0(0) = 0$. It follows that $x_0(t)$ exists and passes through the fold point (see Figure 3.1a). If $x_0(t)$ connects S_a to S_r then, heuristically, one can expect a connection from $S_{a,\varepsilon}$ to $S_{r,\varepsilon}$. In what follows we show that such a connection exists along a curve in the (λ, ε)-plane.

Remark 3.1. The case when $x_0(t)$ connects S_r to S_a can also be treated with the methods of this article, but it is less interesting and will be omitted.

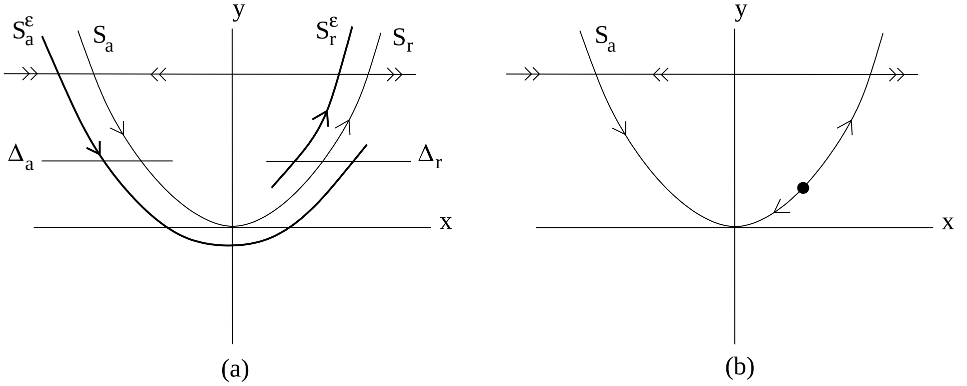


FIG. 3.1. *Reduced flow. (a) $\lambda = 0$, (b) $\lambda > 0$.*

It follows from assumptions (3.2) and (3.3) that using simple coordinate changes (scaling and translations) one can transform (3.1) to the canonical form

$$\begin{aligned} x' &= -yh_1(x, y, \lambda, \varepsilon) + x^2 h_2(x, y, \lambda, \varepsilon) + \varepsilon h_3(x, y, \lambda, \varepsilon), \\ y' &= \varepsilon\left(\pm x h_4(x, y, \lambda, \varepsilon) - \lambda h_5(x, y, \lambda, \varepsilon) + y h_6(x, y, \lambda, \varepsilon)\right), \end{aligned} \tag{3.6}$$

where

$$\begin{aligned} &h_3(x, y, \lambda, \varepsilon) = O(x, y, \lambda, \varepsilon), \\ &h_j(x, y, \lambda, \varepsilon) = 1 + O(x, y, \lambda, \varepsilon), \quad j = 1, 2, 4, 5. \end{aligned}$$

We assume that the sign in front of the term xh_4 is positive. In this case, $x_0(t)$ connects S_a to S_r. Clearly, the signs of the various terms in the above equation correspond to a certain choice of signs in the nondegeneracy conditions made earlier. The parameter λ has been rescaled such that the reduced flow has an equilibrium on S_r for $\lambda > 0$ (see Figure 3.1b).

We introduce the following notation:

$$a_1 = \frac{\partial h_3}{\partial x}(0,0,0,0), \quad a_2 = \frac{\partial h_1}{\partial x}(0,0,0,0), \quad a_3 = \frac{\partial h_2}{\partial x}(0,0,0,0),$$
$$a_4 = \frac{\partial h_4}{\partial x}(0,0,0,0), \quad a_5 = h_6(0,0,0,0),$$

and define

$$A = -a_2 + 3a_3 - 2a_4 - 2a_5. \tag{3.7}$$

The constant A will show up in various computations and results related to the analysis of the dynamics near the canard point. In particular, we will need the genericity condition $A \neq 0$ in the analysis of canard explosion in [14].

For $j = a, r$ let $\Delta_j = \{(x, \rho^2), x \in I_j\}$ be a section of S_j near the fold point with ρ sufficiently small and suitable intervals I_j (see Figure 3.1a). Define $q_{j,\varepsilon} = \Delta_j \cap S_{j,\varepsilon}$. Let π be the transition map for the flow of (3.1) from Δ_a to Δ_r. The following theorem describes the behavior of $S_{a,\varepsilon}$ and $S_{r,\varepsilon}$ near the canard point.

THEOREM 3.1. *Assume that system* (3.1) *satisfies the defining conditions* (3.2)–(3.4) *of a canard point. Assume that the solution $x_0(t)$ of the reduced problem connects S_a to S_r. Then there exists $\varepsilon_0 > 0$ and a smooth function $\lambda_c(\sqrt{\varepsilon})$ defined on* $[0, \varepsilon_0]$ *such that for $\varepsilon \in (0, \varepsilon_0]$ the following assertions hold:*

1. $\pi(q_{a,\varepsilon}) = q_{r,\varepsilon}$ *if and only if* $\lambda = \lambda_c(\sqrt{\varepsilon})$.
2. *The function λ_c has the expansion*

$$\lambda_c(\sqrt{\varepsilon}) = -\left(\frac{a_1 + a_5}{2} + \frac{1}{8}A\right)\varepsilon + O(\varepsilon^{3/2}). \tag{3.8}$$

3. *The transition map π is defined only for λ in an interval around $\lambda_c(\sqrt{\varepsilon})$ of width $O(e^{-c/\varepsilon})$ for some $c > 0$.*
4.

$$\frac{\partial}{\partial \lambda}(\pi(q_{a,\varepsilon}) - q_{r,\varepsilon})\bigg|_{\lambda = \lambda_c(\sqrt{\varepsilon})} > 0.$$

Remark 3.2. For $\lambda = \lambda_c(\sqrt{\varepsilon})$ the slow manifold $S_{a,\varepsilon}$ extends to the slow manifold $S_{r,\varepsilon}$, i.e., the slow manifold consists of a single *canard solution*. We use the term canard solution only for solutions with this property. In other works all solutions of system (1.1) which follow $S_{r,\varepsilon}$ for a time interval of order $O(1)$ are called canard solutions. The solution described in the above theorem is then called a *maximal canard* solution.

For the special case of the van der Pol equation the above results are contained in [6]. Here we treat the general case of a canard point and identify the important parameter A. Also, besides using the same blow-up our proof of the existence of a canard solution based on extending slow manifolds combined with a Melnikov-type argument is new.

3.2. Blow-up. The analysis in this section is, in many aspects, similar to that in section 2. In particular we apply a blow-up transformation, yet the weights (powers of r) must be different than in the case of simple fold. Furthermore, the parameter λ is now included in the blow-up. The blow-up transformation Φ maps $B = S^2 \times [-\mu, \mu] \times [0, \rho]$ to $\mathbb{R}^4$ according to

$$x = \bar{r}\bar{x}, \quad y = \bar{r}^2\bar{y}, \quad \varepsilon = \bar{r}^2\bar{\varepsilon}, \quad \lambda = \bar{r}\bar{\lambda}. \tag{3.9}$$

The constants μ and ρ are chosen small enough such that equations (3.6) are valid in $\Phi(B)$. Let $\bar{X}$ denote the corresponding blown up vector field. In section 2 we used the charts K_1, K_2, and K_3 to obtain the dynamics of $\bar{X}$. Here charts K_1 and K_2 are sufficient to describe the relevant phenomena. In chart K_1, the blow-up transformation (3.9) is

$$x = r_1x_1, \quad y = r_1^2, \quad \varepsilon = r_1^2\varepsilon_1, \quad \lambda = r_1\lambda_1, \tag{3.10}$$

where $(x_1, r_1, \varepsilon_1, \lambda_1)$ are the coordinates in $\mathbb{R}^4$. In chart K_2, the blow-up transformation (3.9) is

$$x = r_2x_2, \quad y = r_2^2y_2, \quad \varepsilon = r_2^2, \quad \lambda = r_2\lambda_2, \tag{3.11}$$

where $(x_2, y_2, r_2, \lambda_2)$ are the coordinates in $\mathbb{R}^4$. A simple computation gives the following lemma.

LEMMA 3.2. *Let κ_{12} denote the change of coordinates from K_1 to K_2. Then κ_{12} is given by*

$$x_2 = x_1\varepsilon_1^{-1/2}, \quad y_2 = \varepsilon_1^{-1}, \quad r_2 = r_1\varepsilon_1^{1/2}, \quad \lambda_2 = \varepsilon_1^{-1/2}\lambda_1 \quad \textit{for } \varepsilon_1 > 0. \tag{3.12}$$

Similarly, $\kappa_{21} = \kappa_{12}^{-1}$ is given by

$$x_1 = x_2y_2^{-1/2}, \quad r_1 = r_2y_2^{1/2}, \quad \varepsilon_1 = y_2^{-1}, \quad \lambda_1 = \lambda_2y_2^{-1/2} \qquad \textit{for } y_2 > 0. \tag{3.13}$$

3.3. Dynamics in chart K_2—preliminary analysis. After dividing out a factor r_2 in a manner analogous to that in section 2, the transformed equations (3.6) have the form

$$\begin{aligned} x_2' &= -y_2 + x_2^2 + r_2G_1(x_2, y_2) + O(r_2(\lambda_2 + r_2)), \\ y_2' &= x_2 - \lambda_2 + r_2G_2(x_2, y_2) + O(r_2(\lambda_2 + r_2)), \end{aligned} \tag{3.14}$$

where

$$G(x_2, y_2) = \begin{pmatrix} G_1(x_2, y_2) \\ G_2(x_2, y_2) \end{pmatrix} = \begin{pmatrix} a_1x_2 - a_2x_2y_2 + a_3x_2^3 \\ a_4x_2^2 + a_5y_2 \end{pmatrix}.$$

Remark 3.3. It turns out that for $r_2 = \lambda_2 = 0$ the system (3.14) is integrable. For this reason the $O(r_2)$ and $O(\lambda_2)$ terms are crucial for the analysis.

Setting $r_2 = \lambda_2 = 0$ in (3.14) we obtain

$$\begin{aligned} x_2' &= -y_2 + x_2^2, \\ y_2' &= x_2. \end{aligned} \tag{3.15}$$

Equation (3.15) is integrable. More precisely, we have the following lemma.

Lemma 3.3. *The function*

$$H(x_2, y_2) = \frac{1}{2}e^{-2y_2}\left(y_2 - x_2^2 + \frac{1}{2}\right) \tag{3.16}$$

is a constant of motion for (3.15).

Proof. A computation gives

$$\begin{aligned} x_2' &= e^{2y_2}\frac{\partial H}{\partial y_2}(x_2, y_2), \\ y_2' &= -e^{2y_2}\frac{\partial H}{\partial x_2}(x_2, y_2). \end{aligned} \tag{3.17}$$

The result follows. □

Lemma 3.3 implies that the solutions of (3.15) are determined by the level curves of $H(x_2, y_2)$. Observe that (3.15) has an equilibrium at $(0,0)$ of center type, surrounded by a family of periodic orbits $H(x_2, y_2) = h$, $h \in (0, 1/4)$. The sets $H(x_2, y_2) = h$, $h \leq 0$, correspond to unbounded solutions. The locus of the solution determined by $h = 0$ is the parabola $x_2^2 - y_2 = 1/2$, and this solution is given by

$$\gamma_{c,2}(t_2) = (x_{c,2}(t_2), y_{c,2}(t_2)) = \left(\frac{1}{2}t_2, \frac{1}{4}t_2^2 - \frac{1}{2}\right), \quad t_2 \in \mathbb{R}.$$

The special solution $\bar{\gamma}_c$ is of central importance to the canard phenomenon. We will see that $\bar{\gamma}_c$ connects the endpoint p_a of the critical manifold S_a across the sphere S^2 to the endpoint p_r of the critical manifold S_r. As in the analysis of the fold point, the points p_a and p_r lie on the equator of S^2 and will be studied in chart K_1. Our goal is to investigate how this connection breaks under perturbation. The tool for this investigation will be a variant of the Melnikov method in which again both charts K_1 and K_2 will be used.

3.4. Dynamics in chart K_1. We proceed in a manner analogous to that in section 2.5 and obtain the following system of equations:

$$\begin{aligned} x_1' &= -1 + x_1^2 + r_1(a_1\varepsilon_1 x_1 - a_2 x_1 + a_3 x_1^3) - \frac{1}{2}\varepsilon_1 x_1 F(x_1, r_1, \varepsilon_1, \lambda_1) \\ &\quad + O(r_1(r_1 + \lambda_1)), \end{aligned} \tag{3.18a}$$

$$r_1' = \frac{1}{2}r_1\varepsilon_1 F(x_1, r_1, \varepsilon_1, \lambda_1), \tag{3.18b}$$

$$\varepsilon_1' = -\varepsilon_1^2 F(x_1, r_1, \varepsilon_1, \lambda_1), \tag{3.18c}$$

$$\lambda_1' = -\frac{1}{2}\lambda_1\varepsilon_1 F(x_1, r_1, \varepsilon_1, \lambda_1), \tag{3.18d}$$

where

$$F(x_1, r_1, \varepsilon_1, \lambda_1) = x_1 - \lambda_1 + r_1(a_4 x_1^2 + a_5) + O(r_1(r_1 + \lambda_1)).$$

It suffices to consider $\lambda_1 \in (-\mu, \mu)$, where $\mu > 0$ can be chosen small. Many features of the dynamics in chart K_1 are analogous to section 2.5; therefore, we provide fewer details. The hyperplanes $r_1 = 0$, $\varepsilon_1 = 0$, and $\lambda_1 = 0$ are invariant, and the invariant line $l_1 := \{(x_1, 0, 0, 0) \; : \; x_1 \in \mathbb{R}\}$ contains two equilibria $p_a = (-1, 0, 0, 0)$ and $p_r = (1, 0, 0, 0)$ which are endpoints of lines of equilibria $S_{a,1}$ and $S_{r,1}$, respectively. For the flow on the line l_1, the equilibrium p_a is attracting and p_r is repelling. Considered as equilibria of system (3.18), both equilibria have a triple eigenvalue zero.

In the invariant plane $r_1 = \lambda_1 = 0$ system (3.18) reduces to

$$\begin{aligned} x_1' &= -1 + x_1^2 - \frac{1}{2}\varepsilon_1 x_1^2, \\ \varepsilon_1' &= -\varepsilon_1^2 x_1. \end{aligned} \tag{3.19}$$

Consequently, the sign of ε_1' is negative for initial conditions near p_r, which implies that the repelling center manifold $N_{r,1}$ at p_r in the half space $\varepsilon_1 > 0$ is unique. The attracting center manifold $N_{a,1}$ at p_a in the half space $\varepsilon_1 > 0$ is also unique, just as in section 2.5. Let

$$D_1 := \{(x_1, r_1, \varepsilon_1, \lambda_1) : \ -2 < x_1 < 2, 0 \le r_1 \le \rho, \ 0 \le \varepsilon_1 \le \delta, -\mu < \lambda_1 < \mu\},$$

where δ, ρ, and μ will be chosen small.

PROPOSITION 3.4. *Choose $c_1 < 2 < c_2$. The constants ρ, δ and μ can be chosen sufficiently small such that the following assertions hold for system* (3.18):

1. *There exists an attracting three-dimensional C^k-center manifold $M_{a,1}$ at p_a that contains the line of equilibria $S_{a,1}$ and the center manifold $N_{a,1}$. In D_1 the manifold $M_{a,1}$ is given as a graph $x_1 = h_a(r_1, \varepsilon_1, \lambda_1)$. The branch of $N_{a,1}$ in $r_1 = \lambda_1 = 0$, $\varepsilon_1 > 0$ is unique and equal to $\gamma_{c,1} := \kappa_{21}(\gamma_{c,2})$, where $\gamma_{c,2}$ is the part of the special trajectory introduced in section* 3.3, *corresponding to x_2 close to $-\infty$.*
2. *There exists a repelling three-dimensional C^k-center manifold $M_{r,1}$ at p_r which contains the line of equilibria $S_{r,1}$ and the center manifold $N_{r,1}$. In D_1 the manifold $M_{r,1}$ is given as a graph $x_1 = h_r(r_1, \varepsilon_1, \lambda_1)$. The branch of $N_{r,1}$ in $r_1 = \lambda_1 = 0$, $\varepsilon_1 > 0$ is unique and equal to $\kappa_{21}(\gamma_{c,2})$ for x_2 close to ∞.*
3. *There exists a stable invariant foliation $\mathcal{F}^s$ with base $M_{a,1}$ and one-dimensional fibers. There exist positive constants $K_{a,1}$ and $K_{a,2}$ such that the contraction along $\mathcal{F}^s$ in a time interval of length T can be estimated by $K_{a,2}e^{-c_2 T}$ from below and by $K_{a,1}e^{-c_1 T}$ from above.*
4. *There exists an unstable invariant foliation $\mathcal{F}^u$ with base $M_{r,1}$ and one-dimensional fibers. There exist positive constants $K_{r,1}$ and $K_{r,2}$ such that the expansion along $\mathcal{F}^u$ in a time interval of length T can be estimated by $K_{r,1}e^{c_1 T}$ from below and by $K_{r,2}e^{c_2 T}$ from above.*

Proof. The proof is analogous to the proof of Proposition 3.4. □

We now define the following sections:

$$\begin{aligned} \Sigma_{a,1}^{in} &:= \{(x_1, r_1, \varepsilon_1, \lambda_1) \in D_1 \ : \ r_1 = \rho, |1 + x_1| < \beta\}, \\ \Sigma_{a,1}^{out} &:= \{(x_1, r_1, \varepsilon_1, \lambda_1) \in D_1 \ : \ \varepsilon_1 = \delta, |1 + x_1| < \beta\}, \\ \Sigma_{r,1}^{in} &:= \{(x_1, r_1, \varepsilon_1, \lambda_1) \in D_1 \ : \ \varepsilon_1 = \delta, |1 - x_1| < \beta\}, \\ \Sigma_{r,1}^{out} &:= \{(x_1, r_1, \varepsilon_1, \lambda_1) \in D_1 \ : \ r_1 = \rho, |1 - x_1| < \beta\}, \end{aligned}$$

where $\beta > 0$ is chosen small. For $j = a, r$, let $\Pi_{j,1}$ be the transition map defined by the flow of (3.18a) from section $\Sigma_{j,1}^{in}$ to $\Sigma_{j,1}^{out}$. With these definitions, results analogous to Lemma 2.7 and Proposition 2.8 hold for the canard point as well.

3.5. Phase portrait on $S_0^{2,+}$. Based on the analysis in charts K_1 and K_2, we can now describe the dynamics of $\bar{X}$ restricted to $S_0^{2,+}$, i.e., for $\bar{r} = \bar{\lambda} = 0$. The equator S^1 is invariant. On S^1 there are four equilibria: p_a, p_r, q_{in}, q_{out}. These equilibria are hyperbolic for the flow on S^1, the points p_a and q_{out} are attracting, and p_r and q_{in} are repelling. The special trajectory $\bar{\gamma}_c$ is a connecting orbit between p_a and p_r. Besides $\bar{\gamma}_c$, there are three types of orbits in $S^{2,+}$: a concentric family of periodic orbits, an equilibrium of center type, and a family of orbits joining q_{in} to q_{out}. The corresponding phase portrait is shown in Figure 3.2.

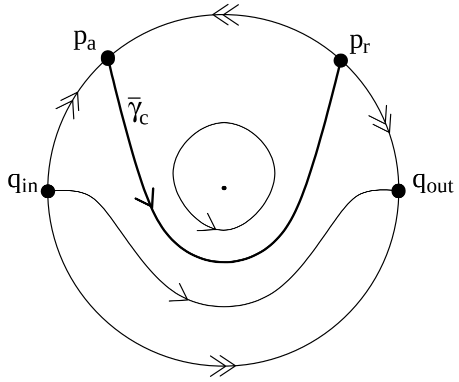


FIG. 3.2. *Blow-up of a canard point restricted to $S_0^{2,+}$ for $\bar{\lambda} = 0$.*

Remark 3.4. It turns out that the connection $\bar{\gamma}_c$ from p_a to p_r breaks for $\bar{\lambda} \neq 0$. Some important aspects of the dynamics near $S_0^{2,+}$ leading to Theorem 3.1 can be understood by investigating whether the intersection of the three-dimensional center manifolds $\bar{M}_a$ and $\bar{M}_r$ at $\bar{\gamma}_c$ is transverse. The relevant computation is carried out in the next section. Bifurcations from the center equilibrium and the family of periodic orbits and their relevance to canard explosion are studied in [14].

3.6. Melnikov computation of the separation between $M_{a,2}$ and $M_{r,2}$. The analysis of the previous sections implies that $M_{a,2}$ and $M_{r,2}$ intersect along $\gamma_{c,2}$ for $r_2 = \lambda_2 = 0$. To prove Theorem 3.1 we compute the first order separation between $M_{a,2}$ and $M_{r,2}$ with respect to r_2 and λ_2. The splitting of the manifolds off the sphere for $\lambda_2 = 0$ is shown in Figure 3.3. All functions defined below are considered as functions of $r_2 \in [0, \rho]$ and $\lambda_2 \in (-\mu, \mu)$, for small $\rho > 0$ and $\mu > 0$, without indicating this dependence explicitly. In the formulas below we drop the subscript of the time variable t_2 from chart K_2, i.e., $t = t_2$.

Let $\gamma_{a,1}$ be the trajectory of system (3.18) contained in $M_{a,1}$ for which $r_1\sqrt{\varepsilon_1} = r_2$. Let $\gamma_{a,2}(t) = (x_{a,2}(t), y_{a,2}(t))$ be the continuation of $\gamma_{a,1}$ to chart K_2, i.e., $\gamma_{a,2}$ is a solution of (3.14), parametrized such that $x_{a,2}(0) = 0$. Analogously, let $\gamma_{r,1}$ be the trajectory of (3.18) contained in $M_{r,1}$ for which $r_1\sqrt{\varepsilon_1} = r_2$ and let $\gamma_{r,2}(t) = (x_{r,2}(t), y_{r,2}(t))$ be the corresponding, backward continued solution of (3.14) parametrized such that $x_{r,2}(0) = 0$.

With these definitions, measuring the separation of $M_{a,2}$ and $M_{r,2}$ corresponds to measuring $y_{a,2}(0) - y_{r,2}(0)$, which is equivalent to estimating the *distance function*

$$\mathcal{D}_c(r_2, \lambda_2) := H(0, y_{a,2}(0)) - H(0, y_{r,2}(0)), \tag{3.20}$$

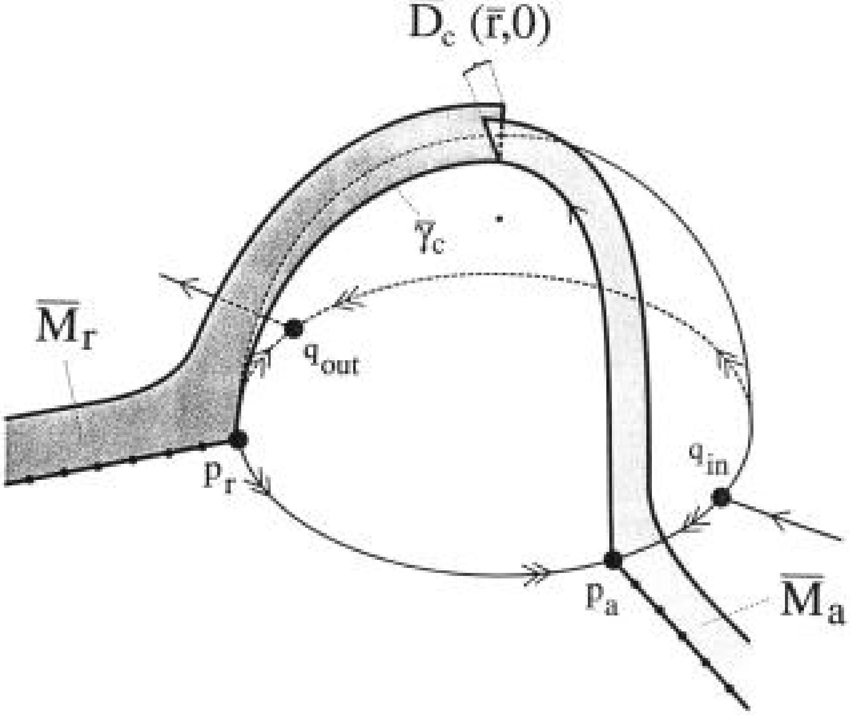


FIG. 3.3. *Splitting of $\bar{M}_a$ and $\bar{M}_r$ for $\bar{\lambda} = 0$.*

since $\frac{\partial H}{\partial y_2}(0, y_2) \neq 0$ for $y_2 < 0$. We define

$$d_{r_2} = \int_{-\infty}^{\infty} \text{grad}H(\gamma_{c,2}(t)) \cdot G(\gamma_{c,2}(t))dt, \tag{3.21}$$

where G is the function defined in section 3.3. Similarly we define

$$d_{\lambda_2} = \int_{-\infty}^{\infty} \text{grad}H(\gamma_{c,2}(t)) \cdot \begin{pmatrix} 0 \\ -1 \end{pmatrix} dt. \tag{3.22}$$

We will prove the following result.

PROPOSITION 3.5. *For ρ and μ small enough, the distance function $\mathcal{D}_c(r_2, \lambda_2)$ is a C^k-function and has the expansion*

$$\mathcal{D}_c(r_2, \lambda_2) = d_{r_2} r_2 + d_{\lambda_2}\lambda_2 + O(2). \tag{3.23}$$

Proof. By construction $\mathcal{D}_c$ is C^k smooth and $\mathcal{D}_c(0,0) = 0$. Thus we have to verify the expansion. We carry out the computation for r_2. The result for λ_2 can be obtained in a similar way. We set $\lambda_2 = 0$ and consider $r_2 \in [0, \rho]$. We will show that

$$H(0, y_{a,2}(0)) = r_2 \int_{-\infty}^{0} \text{grad}H(\gamma_{c,2}(t)) \cdot G(\gamma_{c,2}(t))dt + O(r_2^2). \tag{3.24}$$

An analogous argument yields

$$H(0, y_{r,2}(0)) = -r_2 \int_{0}^{\infty} \text{grad}H(\gamma_{c,2}(t)) \cdot G(\gamma_{c,2}(t))dt + O(r_2^2), \tag{3.25}$$

which implies the proposition.

We define $T(r_2, \delta) < 0$ such that $y_{a,2}(T) = \delta^{-1}$, where δ is the constant from the definition of $\Sigma_{1,a}^{out}$. We write

$$H(0, y_{a,2}(0)) = H(x_{a,2}(T), \delta^{-1}) + \int_T^0 \frac{dH}{dt}(\gamma_{a,2}(t))dt.$$

By standard methods [4]

$$\int_T^0 \frac{dH}{dt}(\gamma_{a,2}(t))dt = r_2 \int_T^0 \operatorname{grad} H(\gamma_{a,2}(t)) \cdot G(\gamma_{a,2}(t))dt + O(r_2^2).$$

It turns out to be very natural to compute $H(x_{a,2}(T), \delta^{-1})$ in chart K_1. We begin by parametrizing $\gamma_{a,1}$ by ε_1, i.e.,

$$\gamma_{a,1}(\varepsilon_1) = \left(x_{a,1}(\varepsilon_1), \frac{r_2}{\sqrt{\varepsilon_1}}, \varepsilon_1, 0\right), \quad \varepsilon_1 \in \left[\left(\frac{r_2}{\rho}\right)^2, \delta\right],$$

where ρ is the constant used in the definition of $\Sigma_{a,1}^{in}$. Let $H_1 = H \circ \kappa_{12}$, i.e.,

$$H_1(x_1, \varepsilon_1) = H\left(\frac{x_1}{\sqrt{\varepsilon_1}}, \frac{1}{\varepsilon_1}\right) = e^{-2/\varepsilon_1}\left(\frac{1}{4} + \frac{1}{2\varepsilon_1} - \frac{x_1^2}{2\varepsilon_1}\right). \tag{3.26}$$

We wish to estimate $H_1(x_{a,1}(\delta), \delta)$. Note that $H_1(x_{a,1}((\frac{r_2}{\rho})^2), (\frac{r_2}{\rho})^2)$ is exponentially small in r_2. Hence,

$$H_1(x_{a,1}(\delta), \delta) = \int_{(\frac{r_2}{\rho})^2}^{\delta} \frac{d}{d\varepsilon_1} H_1(x_{a,1}(\varepsilon_1), \varepsilon_1)d\varepsilon_1 + O(r_2^2).$$

From (3.26) we obtain

$$\begin{aligned} \frac{\partial H_1}{\partial x_1} &= -e^{-2/\varepsilon_1}\frac{x_1}{\varepsilon_1}, \\ \frac{\partial H_1}{\partial \varepsilon_1} &= e^{-2/\varepsilon_1}\frac{1}{\varepsilon_1^2}\left(\frac{1}{2}x_1^2 - \frac{x_1^2}{\varepsilon_1} + \frac{1}{\varepsilon_1}\right). \end{aligned} \tag{3.27}$$

Note that $\frac{dH_1}{d\varepsilon_1}$ evaluated along a trajectory of (3.18) is given by

$$\frac{dH_1}{d\varepsilon_1} = \frac{\partial H_1}{\partial x_1}\frac{x_1'}{\varepsilon_1'} + \frac{\partial H_1}{\partial \varepsilon_1},$$

where x_1' and ε_1' are given by (3.18a) and (3.18c), respectively. By using (3.27), expanding and using the relation $r_1 = r_2/\sqrt{\varepsilon_1}$ to eliminate r_1, we obtain the following formula for $\frac{dH_1}{d\varepsilon_1}$ evaluated along a trajectory of (3.18):

(3.28)
$$\frac{dH_1}{d\varepsilon_1}(x_1, \varepsilon_1) = e^{-2/\varepsilon_1}\varepsilon_1^{-7/2}\left[r_2(a_1x_1\varepsilon_1 - a_2x_1 + a_3x_1^3 + \frac{1 - x_1^2}{x_1}(a_4x_1^2 + a_5)) + \frac{1}{\sqrt{\varepsilon_1}}O(r_2^2)\right].$$

We set

$$\eta(x_1, \varepsilon_1) = e^{-2/\varepsilon_1}\varepsilon_1^{-7/2}\left(a_1x_1\varepsilon_1 - a_2x_1 + a_3x_1^3 + \frac{1 - x_1^2}{x_1}(a_4x_1^2 + a_5)\right). \tag{3.29}$$

It follows that

$$\int_{(\frac{r_2}{\rho})^2}^{\delta} \frac{d}{d\varepsilon_1} H_1(x_{a,1}(\varepsilon_1),\varepsilon_1)d\varepsilon_1 = r_2 \int_{(\frac{r_2}{\rho})^2}^{\delta} \eta(x_{a,1}(\varepsilon_1),\varepsilon_1)d\varepsilon_1 + O(r_2^2),$$

where we have used that the integral of the error term in (3.28) is $O(r_2^2)$ because of the exponentially small prefactor. To complete the computation, recall that $(x_{c,1}(\varepsilon_1),\varepsilon_1)$ parametrizes $N_{a,1}$, and, by center manifold theory,

$$|x_{c,1}(\varepsilon_1) - x_{a,1}(\varepsilon_1)| = O(r_1) = O\left(\frac{r_2}{\sqrt{\varepsilon_1}}\right).$$

By using this estimate it follows that

$$H_1(x_{a,1}(\delta),\delta) = r_2 \int_0^{\delta} \eta(x_{c,1}(\varepsilon_1),\varepsilon_1)d\varepsilon_1 + O(r_2^2),$$

where we have again used the exponentially small prefactor of the integrand to estimate the error caused by first replacing $x_{a,1}$ by $x_{c,1}$ and then changing the interval of integration to $[0,\delta]$. By applying the change of variables formula, we transform this integral to chart K_2 and obtain

$$\int_0^{\delta} \eta(x_{c,1}(\varepsilon_1),\varepsilon_1)d\varepsilon_1 = \int_{-\infty}^{T} \mathrm{grad} H(\gamma_{c,2}(t)) \cdot G(\gamma_{c,2}(t))dt.$$

The result follows. □

Remark 3.5. The formulas for the constants d_{r_2} and d_{λ_2} are the usual Melnikov integrals for the splitting of saddle-saddle connections for perturbations of planar Hamiltonian vector fields. However, the situation considered above is not covered by the usual Melnikov theory.

Proof of Theorem 3.1. It follows from the above results that for $(r_2,\lambda_2) \in [0,\rho) \times (-\mu,\mu)$ a connection from $S_{a,\varepsilon}$ to $S_{r,\varepsilon}$ exists if and only if

$$\mathcal{D}_c(r_2,\lambda_2) = 0. \tag{3.30}$$

We have shown that $\mathcal{D}_c(r_2,\lambda_2) = d_{r_2}r_2 + d_{\lambda_2}\lambda_2 + O(2)$. Hence, (3.30) can be solved for λ_2 by the implicit function theorem provided that $d_{\lambda_2} \neq 0$. The solution has the expansion

$$\lambda_2 = -\frac{d_{r_2}}{d_{\lambda_2}} r_2 + O(r_2^2). \tag{3.31}$$

By using the parametrization of $\gamma_{c,2}$ and repeated integration by parts, we compute

$$\begin{aligned} d_{r_2} &= \int_{-\infty}^{\infty} e^{-2y_2}(-a_1x_2^2 + (a_2 - a_4 + a_5)x_2^2y_2 + (a_4 - a_3)x_2^4 - a_5y_2^2)dt \\ &= -\frac{e}{4}\left(a_1 + a_5 + \frac{1}{4}A\right)\int_{-\infty}^{\infty} e^{-t^2/2}dt \end{aligned} \tag{3.32}$$

and

$$d_{\lambda_2} = -\int_{-\infty}^{\infty} \frac{1}{2}e^{-2y_2}dt = -\frac{e}{2}\int_{-\infty}^{\infty} e^{-t^2/2}dt. \tag{3.33}$$

This proves assertion (1). Property (2) follows by applying transformation (3.11) to the the expansion (3.31). Assertion (3) follows from the above combined with Proposition 3.4 because $M_{a,2}$ must be exponentially close to $M_{r,2}$ to reach the section $\Sigma_{r,1}^{out}$. Finally, the inequality $d_{\lambda_2} < 0$ implies that the intersection of the slow manifolds breaks as described in assertion (4). □

Remark 3.6. It is known that the function λ_c from Theorem 3.1 which describes the canard curve actually has a power series in ε; see [3], [7], [19]. It is interesting to observe that this fact can be explained by a symmetry property of the blow-up transformation. The form of the blow-up transformation (3.9) implies that the transformation

$$(\bar{x}, \bar{y}, \bar{r}, \bar{\lambda}) \mapsto (-\bar{x}, \bar{y}, -\bar{r}, -\bar{\lambda}) \tag{3.34}$$

is a time-reversal symmetry of the blown-up vector field $\bar{X}$; i.e., it maps orbits to orbits. Note that for prescribed slow manifolds outside V, the canard curve is uniquely determined. Since the transformation (3.34) maps a canard curve to a canard curve, it follows that $-\lambda_{c,2}(r_2) = \lambda_{c,2}(-r_2)$ and consequently $\lambda_c(r_2) = \lambda_{c,2}(-r_2)$. From this we conclude that an asymptotic expansion of the canard curve is quadratic in r_2 and hence is a power series in ε. If the original vector field is smooth and the slow manifolds are chosen smooth, then, by a theorem of Schwartz [21], λ_c is a smooth function of ε.

The function λ_c depends on the choice of the slow manifolds. However all slow manifolds are exponentially close and the corresponding functions λ_c differ only by exponentially small terms and have the same asymptotic expansion.

Acknowledgments. The authors would like to thank Freddy Dumortier, Stephan van Gils, Alexandra Milik, and Martin Wechselberger for helpful discussions.

The authors also acknowledge the support and hospitality of the Institute for Mathematics and Its Applications at the University of Minnesota, where a part of this research was carried out.

REFERENCES

[1] V. I. Arnold, ed., *Dynamical Systems* V, *Bifurcation theory and catastrophe theory,* Encyclopaedia Math. Sci. 5, Springer-Verlag, Berlin, 1989.

[2] E. Benoit, ed., *Dynamic Bifurcations*, Lecture Notes in Math. 1493, Springer, New York, 1991.

[3] E. Benoit, J. L. Callot, F. Diener, and M. Diener, *Chasse au canards*, Collect. Math., 31 (1981), pp. 37–119.

[4] S-N. Chow, C. Li, and D. Wang, *Normal Forms and Bifurcation of Planar Vector Fields*, Cambridge University Press, Cambridge, 1994.

[5] F. Dumortier, *Techniques in the theory of local bifurcations: Blow-up, normal forms, nilpotent bifurcations, singular perturbations*, in Bifurcations and Periodic Orbits of Vector Fields, D. Szlomiuk, ed., NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 408, Kluwer, Dordrecht, the Netherlands, 1993, pp. 19–73.

[6] F. Dumortier and R. Roussarie, *Canard cycles and center manifolds*, Mem. Amer. Math. Soc. 577, Providence, 1996.

[7] W. Eckhaus, *Relaxation oscillations including a standard chase on French ducks*, in Asymptotic Analysis II, Lecture Notes in Math. 985, Springer, New York, 1983, pp. 449–494.

[8] N. Fenichel, *Geometric singular perturbation theory*, J. Differential Equations, 31 (1979), pp. 53–98.

[9] J. Grasman, *Asymptotic Methods for Relaxation Oscillations and Applications*, Springer, New York, 1987.

[10] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer, New York, 1983.

[11] C. K. R. T. Jones, *Geometric singular perturbation theory*, in Dynamical Systems, Lecture Notes in Math. 1609, Springer, New York, 1995, pp. 44–120.

[12] J. Kevorkian and J. D. Cole, *Perturbation Methods in Applied Mathematics*, Springer, New York, 1981.

[13] M. Krupa and P. Szmolyan, *Geometric analysis of the singularly perturbed planar fold, Multiple time scale dynamical systems*, IMA Vol. Math. Appl., 122 (2000), pp. 89–116.

[14] M. Krupa and P. Szmolyan, *Relaxation oscillation and canard explosion*, J. Differential Equations, to appear, 2001.

[15] M. Krupa and P. Szmolyan, *Transcritical and pitchfork singularities of critical manifolds*, Nonlinearity, submitted, 2000.

[16] S.A. van Gils, M. Krupa, and P. Szmolyan, *Asymptotic expansions using blow-up*, Z. Angew. Math. Phys., submitted, 2000.

[17] A. Milik and P. Szmolyan, *Multiple time scales and canards in a chemical oscillator*, *Multiple time scale dynamical systems*, IMA Vol. Math. Appl., 122 (2000), pp. 117–140.

[18] E. F. Mishchenko and N. Kh. Rozov, *Differential Equations with Small Parameters and Relaxation Oscillations*, Plenum Press, New York, 1980.

[19] E. F. Mishchenko, Yu. S. Kolesov, A. Yu. Kolesov, and N. Kh. Rozov, *Asymptotic Methods in Singularly Perturbed Systems*, Monogr. Contemp. Math., Consultants Bureau, New York, 1994.

[20] L. S. Pontryagin, *Asymptotic behavior of solutions of systems of differential equations with a small parameter in derivatives of highest order*, Izv. Akad. Nauk. SSSR Ser. Mat., 21 (1957), pp. 605–626.

[21] G. Schwartz, *Smooth functions invariant under the action of a compact Lie group*, Topology, 14 (1975), pp. 63–68.

[22] S. Sternberg, *On the nature of local homeomorphisms of Euclidean n-space* II, Amer. J. Math., 80 (1958), pp. 623–631.

[23] M. Wechselberger, *Singularly Perturbed Folds and Canards in* $\mathbb{R}^3$, Ph.D. Thesis, TU-Wien, Wien, Austria, 1998.

SIAM J. MATH. ANAL.
Vol. 33, No. 2, pp. 315–346

INFINITE DIMENSIONAL GEOMETRIC SINGULAR PERTURBATION THEORY FOR THE MAXWELL–BLOCH EQUATIONS*

GOVIND MENON[†] AND GYÖRGY HALLER[‡]

Abstract. We study the Maxwell–Bloch equations governing a two-level laser in a ring cavity. For Class A lasers, these equations have two widely separated time scales and form a singularly perturbed, semilinear hyperbolic system with two distinct characteristics. We extend Fenichel's geometric singular perturbation theory [N. Fenichel, *J. Differential Equations*, 31 (1979), pp. 53–98] to the Maxwell–Bloch equations by proving the persistence of a $C^k, 0 < k < \infty$, slow manifold under an unbounded perturbation. The proof is obtained by a modified graph transform method. We use uniform decay estimates of Constantin, Foias, and Gibbon [*Nonlinearity*, 2 (1989), pp. 241–269] to obtain a cone condition. These estimates rely on the energy preserving nature of the nonlinearity and the existence of two distinct characteristics. The cone condition and the fact that the unbounded perturbation generates a continuous group are used to define the graph transform. The slow manifold is a globally attracting, positively invariant manifold, with infinite dimension and codimension, that contains the attractor of the system. The slow manifold depends only continuously on ε and converges uniformly on (strongly) compact sets to the critical manifold. This enables us to rigorously decouple the slow and fast time scales and obtain a reduced (but still infinite-dimensional) dynamical system described by a functional differential equation.

Key words. Maxwell–Bloch equations, invariant manifolds, geometric singular perturbation theory

AMS subject classifications. 58F30, 35Q, 78A60, 37L, 35B

PII. S0036141000360458

1. Introduction.

1.1. The Maxwell–Bloch equations.

We shall study the asymptotic dynamics of the laser equations proposed by Risken and Nummedal [25]. These are amplitude equations describing a two-level laser derived by a semiclassical approximation. The electric field obeys the classical Maxwell equations, and the light-matter interaction is modeled by the quantum mechanical Bloch equations. There are numerous simplifications in this model, several of which are pointed out in [25]. Nevertheless, these equations are quite faithful to the underlying physics and are also mathematically tractable in certain limits. The equations we will study are

$$E_\tau + E_x = \kappa(P - E), \tag{1.1}$$

$$P_\tau = \gamma_\perp[ED - (1 + i\delta)P], \tag{1.2}$$

$$D_\tau = \gamma_\parallel \left[\lambda + 1 - D - \frac{\lambda}{2}(E^* P + E P^*) \right]. \tag{1.3}$$

$E, P \in \mathbb{C}$, and $D \in \mathbb{R}$ are periodic on the domain $[0, 1]$; E is the electric field, P is the polarization of the gain medium, and D is a measure of the population inversion;

*Received by the editors March 18, 2000; accepted for publication (in revised form) November 2, 2000; published electronically July 19, 2001.
http://www.siam.org/journals/sima/33-2/36045.html

[†] Max-Planck-Institute for Mathematics in the Sciences, Inselstr. 22-26, 04103 Leipzig, Germany (govind.menon@mis.mpg.de). This author was supported by a Wonderlic Fellowship from Brown University.

[‡] Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Box F, Brown University, Providence, RI 02912 (haller@cfm.brown.edu). This author was partially supported by NSF grant DMS-98-00922 and an Alfred. P. Sloan Fellowship.

$\kappa, \gamma_{\perp}, \gamma_{\|} > 0$ are phenomenological damping constants; $\lambda > 0$ is a pumping term; $\delta \in \mathbb{R}$ is a detuning parameter. All the variables are dimensionless and have been scaled to the continuous wave (cw) solutions. By these we mean spatially homogeneous steady states of (1.1)–(1.3) that correspond to a steady output from the laser. In another scaling these equations are also known as the Lorenz PDE. The Lorenz ODEs are contained in the system (1.1)–(1.3) when $\delta = 0$, and attention is restricted to real valued, spatially independent solutions.

Constantin, Foias, and Gibbon were the first to study these equations rigorously [9]. They proved the existence of global weak solutions and a C^{∞} global attractor in L^2 with finite Hausdorff dimension. Recently Xin and Moloney studied the equations in three dimensions with the addition of a transverse dispersive term [28]. They proved the existence and uniqueness of weak solutions in L^p, $2 \leq p < \infty$, and an attractor with finite regularity. Naturally, one must expect the dynamics to depend strongly on the parameter values, and in some parameter ranges the analysis will be easier than in others. Kovacic and Wettergren studied the Maxwell–Bloch equations (in a different scaling) near an integrable limit in [20, 27], respectively. The motivation there is to use the knowledge of the geometry of the integrable limit to understand the dynamics when the damping and forcing are small.

1.2. Adiabatic elimination for Class A lasers. Different types of lasers have vastly different dynamics because of the wide variation in parameters $\kappa, \gamma_{\perp}$, and $\gamma_{\|}$. Arecchi proposed a characterization of lasers based on the range of damping parameters, and we shall consider what he terms Class A lasers [1, p. 17]. For this class of lasers, we have $\gamma_{\perp} \approx \gamma_{\|} \gg \kappa$. This scaling has also been called the good cavity limit [10], but strictly speaking, the good cavity limit refers to the case where $\gamma_{\perp} + \gamma_{\|} \geq \kappa$ and includes a much broader range of dynamics than we consider. Nevertheless, the range of Class A lasers is sufficiently wide to be physically and mathematically interesting. For Class B lasers, $\gamma_{\perp} \gg \kappa \tilde{>} \gamma_{\|}$, and for Class C lasers, all three damping constants are comparable. For Class A and B lasers one may hope to simplify the dynamics by separating the evolution on fast and slow time scales. Such adiabatic eliminations are common in the physics literature (see the expository article [1] and the references therein). Our aim is to examine such a reduction from a more mathematical viewpoint. The simplest case is of Class A lasers, and we consider only this scaling henceforth.

Let $\gamma_{\perp} \rightarrow \frac{1}{\varepsilon}, \gamma_{\|} \rightarrow \frac{\gamma_{\|}}{\varepsilon}$, with $0 < \varepsilon \ll 1$. The laser equations are rewritten as

$$E_{\tau} + E_x = \kappa(P - E), \tag{1.4}$$

$$\varepsilon P_{\tau} = ED - (1 + i\delta)P, \tag{1.5}$$

$$\varepsilon D_{\tau} = \gamma_{\|} \left[\lambda + 1 - D - \frac{\lambda}{2}(E^* P + E P^*) \right]. \tag{1.6}$$

This suggests we formally eliminate the "fast" variables P and D, i.e., we set the left-hand sides of (1.5) and (1.6) to zero, solve for P, D as functions of E, and substitute in (1.4). This adiabatic approximation is often used in the physics literature [1, 10, 15] although it is typically used with finite-dimensional modal truncations [16, pp. 156, 290]. It is not apparent that this formal reduction is valid or if the solutions to the singular limit are similar to those of the full system. Indeed, we will show that this reduction leads to false predictions about the asymptotic behavior.

The failure of the formal reduction should not be unexpected. The laser equations are a semilinear hyperbolic system with two characteristics: $x - t$ =constant and

x =constant. The formal reduction procedure eliminates one of these characteristics and thus neglects essential information. Nevertheless, there is some merit in studying the reduced system since it provides some insight into the range of possible asymptotic behavior. We may then attempt to verify if similar behavior persists when $\gamma_\perp$ and $\gamma_\|$ are sufficiently large but finite.

1.3. Geometric singular perturbation theory. A rigorous geometric theory for singularly perturbed ODE was developed by Fenichel [12]. To apply his methods to this problem, one would proceed as follows. First, one regularizes the problem by rescaling time, $t = \frac{\tau}{\varepsilon}$. The new time scale, t, is referred to as fast time. In this variable, the laser equations are

$$E_t = \varepsilon[-E_x + \kappa(P - E)], \tag{1.7}$$

$$P_t = ED - (1 + i\delta)P, \tag{1.8}$$

$$D_t = \gamma_\| \left[\lambda + 1 - D - \frac{\lambda}{2}(E^* P + EP^*) \right]. \tag{1.9}$$

Here E changes slowly with time ($O(\varepsilon)$), and P and D have a time rate of change that is $O(1)$. In the limit $\varepsilon = 0$ the slow variable E is constant. The fast variables still change rapidly except at the equilibria of (1.7)–(1.9). Solving for these equilibria we see that they form a manifold, $\mathcal{M}_0$, given as a graph over the slow variable E. Thus the singularities of the slow time system (1.4)–(1.6) are equilibria of the fast time system (1.7)–(1.9). The formal reduction is equivalent to the assumption that $\mathcal{M}_0$ remains invariant and there is a well-defined flow in slow time restricted to it. How good is this assumption? Some intuition is provided by considering ODE.

The underlying geometry is essentially the same for singularly perturbed ODE. We are given a manifold of equilibria and we want to justify a reduction of the flow to this manifold. Under the crucial hypothesis of *normal hyperbolicity*, Fenichel proved that a compact manifold of equilibria, $\mathcal{M}_0$, continues smoothly to a family of *slow manifolds*, $\mathcal{M}_\varepsilon$, for sufficiently small $\varepsilon > 0$. Furthermore, if we consider the augmented system obtained by appending the equation $\varepsilon_t = 0$ to the ODE, then these manifolds are contained in a global center manifold given as a graph over the slow variable and ε. The singular perturbation problem is then reduced to a regular perturbation problem restricted to this center manifold, and asymptotic expansions in ε are reduced to Taylor series calculations. Fenichel's methods are powerful, and several problems that lie outside the reach of conventional (and typically heuristic) asymptotic methods are easily studied within his framework. There has been much progress in this area; see [19] for a readable introduction.

For PDE the situation is not so simple. There are several obstructions, some of which are technical; for instance, the phase space is no longer locally compact. But another obstruction is essential. For $\varepsilon > 0$ the perturbed flow is not close to the unperturbed flow in the C^1 topology because of the unbounded operator $\varepsilon\partial_x$. Hence there are no general persistence theorems that one can invoke to prove the existence of the slow manifolds $\mathcal{M}_\varepsilon$ (the definitive results in this direction are due to Bates, Lu, and Zeng and may be found in a set of articles beginning with [4]). Furthermore, even if these manifolds exist, one should not expect them to fit together smoothly in ε or to be contained in a smooth global center manifold restricted to which we obtain a regular perturbation problem. Thus there are difficulties in justifying the existence of asymptotic expansions. The addition of an unbounded perturbation also leads to some unexpected phenomenon. For example, one finds new instabilities that are hidden in the adiabatic elimination. Risken and Nummedal [25] showed that the cw solutions of (1.1)–(1.3) are linearly unstable for sufficiently large λ. On the other hand, the adiabatic elimination predicts that these solutions are always stable. We comment on this point again in section 6.

1.4. Main results. The goals of this paper are to understand rigorously the relation between the adiabatic elimination and the full Maxwell–Bloch system and to develop in the process geometric singular perturbation theory for PDE in the setting of a concrete example. Our main theorem, Theorem 4.1, establishes the persistence of a globally attracting, positively invariant manifold diffeomorphic to the manifold of equilibria. This manifold contains the attractor of the system. In infinite dimensions, the persistence of a global invariant manifold under unbounded perturbations is itself a significant fact, and Theorem 4.1 lies considerably outside the scope of general theorems in this field (see, e.g., [2, 3, 8, 4]). This being said, we must in fairness note that the proof relies strongly on the structure of the Maxwell–Bloch equations and is special to this system. There are two facts that play a key role in the analysis. The first is that the addition of the unbounded perturbation for $\varepsilon > 0$ corresponds to the splitting of characteristics that are parallel in the limit. The second is that the nonlinearities of the Maxwell–Bloch equations satisfy strong energy estimates that follow from their physical origin. In particular, the nonlinearities in (1.1)–(1.3) appear only as skew terms and ensure the uniform decay of the polarization and inversion (see 2.5). These estimates were derived by Constantin, Foias, and Gibbon [9]. We utilize these estimates to establish a cone condition of the flow similar to that in [4, 13]. The cone condition, and the fact that the unbounded perturbation generates a continuous group, are crucial ingredients of the proof.

The convergence of the slow manifold to the critical manifold is subtly altered by the unbounded perturbation. We are only able to prove that the convergence is uniform on strongly compact sets (Theorem 6.1). However, one should keep in mind that the phase space is not locally compact.

As a sidelight, we note that the persistence theorem provides an example of an inertial manifold (albeit infinite dimensional) in a problem with no diffusion. Infinite-dimensional inertial manifolds for reaction diffusion equations coupled to ODE (e.g., the Hodgkin–Huxley equations) have been studied by Marion [22]. Marion's methods are a natural complement to methods used for reaction diffusion equations [13] and depend on the control over high wave numbers provided by diffusion. Our methods are quite different and depend strongly on the absence of diffusion.

The rest of this paper is organized as follows. Section 2 contains a priori estimates and results on well-posedness. Section 3 studies the peculiarities of the singular limit. Sections 4 and 5 are dedicated to a proof of the main theorem. The existence of the invariant manifold provides a basis for rigorously decoupling the slow and fast time scales in the system. This is considered in section 6. We also remark on the relation between the formal limit and the slow dynamics there.

2. Existence and uniqueness.

In this section we will prove that the laser equations define a smooth (C^∞) dynamical system in the space of continuous functions. Constantin, Foias, and Gibbon [9] proved that the laser equations define a Lipschitz dynamical system in L^2. The reason for choosing a more restrictive phase space is that smoothness of the flow is essential for invariant manifold techniques. The obstruction to smoothness in L^2 is the quadratic nonlinearity in (1.8) and (1.9). The product of two L^2 functions does not lie in L^2 in general. For continuous functions, however, multiplication is a smooth map. The motivation for choosing L^2 as a phase space is that the laser rises out of noise and the initial data cannot be prepared to be smooth. In view of this, choosing C^0 as the phase space is obviously a restriction in our study. Nevertheless, Theorem 4.2 of [9] states that asymptotically all solutions approach an attractor composed of C^∞ functions. Thus, in order to study asymptotic behavior it is sufficient to restrict our attention to continuous functions.

Our work relies strongly on the a priori estimates proved by Constantin, Foias, and Gibbon and the estimates of this section are largely rescaled versions of their work [9]. A derivation of these estimates, motivated by the underlying physics, may be found in their work. We make no claim to originality for these estimates, and they are included for completeness and to prove well-posedness of the laser equations in a form appropriate for this paper.

To better illustrate the structure of the equations, we rescale the dependent variables. Define $\mu = \sqrt{\lambda\gamma_{\|}}$ and set

$$u = E\ , \quad v = \mu P, \ w = D. \tag{2.1}$$

Thus, (1.7)–(1.9) are transformed to

$$u_t = \varepsilon\left[-u_x + \kappa\left(\frac{v}{\mu} - u\right)\right], \tag{2.2}$$

$$v_t = \mu u w - (1 + i\delta)v, \tag{2.3}$$

$$w_t = \gamma_{\|}(\lambda + 1 - w) - \frac{\mu}{2}(u^* v + u v^*). \tag{2.4}$$

2.1. Notation. The space of continuous functions from the circle into a Euclidean space $\mathbb{E}$ is denoted by $C(S^1;\mathbb{E})$. The phase space for our dynamical system is $\mathbb{X} = \mathbb{X}_1 \times \mathbb{X}_2$, where $\mathbb{X}_1 = C(S^1;\mathbb{C})$ and $\mathbb{X}_2 = C(S^1;\mathbb{C}\times\mathbb{R})$. A typical element of $\mathbb{X}$ is denoted by the triplet (u, v, w). The norm in $\mathbb{X}_1$ is $\|u\| = \sup_{x\in S^1}|u(x)|$, and the norm of $(v, w) \in \mathbb{X}_2$ is $\|(v, w)\| = \sup_{x\in S^1}(|v(x)|^2 + |w(x)|^2)^{1/2}$. The norm of $(u, v, w) \in \mathbb{X}$ is $(\|u\|^2 + \|(v, w)\|^2)^{1/2}$. The projections from $\mathbb{X}$ into $\mathbb{X}_i$ are denoted by Π_i. The space of k-linear maps between two Banach spaces $\mathbb{Y}_1$ and $\mathbb{Y}_2$ will be denoted as $L^k(\mathbb{Y}_1, \mathbb{Y}_2)$. For $k = 1$, we drop the superscript.

2.2. A priori estimates. Notice that if $\varepsilon > 0$, it is sufficient to obtain a priori estimates for either the slow or fast system since they are equivalent. In the rest of this section $\varepsilon > 0$ is fixed.

We first derive a pointwise decay estimate. For all $x \in S^1$, we have

$$\begin{aligned}\partial_t(|v(t,x)|^2 + |w(t,x)|^2) &= -2|v|^2 - 2\gamma_{\|}|w|^2 + 2(\lambda+1)\gamma_{\|}w\\ &\le -2\beta(|v|^2 + |w|^2) + \gamma_{\|}(\lambda+1)^2,\end{aligned}$$

where $\beta = \min(1, \gamma_{\|}/2)$. Integrating the resulting inequality and taking the sup over $x \in S^1$ we obtain

$$\begin{aligned}\|(v,w)(t)\|^2 &\le e^{-2\beta t}(\|(v,w)(0)\|^2) + (\lambda+1)^2\frac{\gamma_{\|}}{2\beta}|1 - e^{-2\beta t}|\\ &=: e^{-2\beta t}(\|(v,w)(0)\|^2) + \rho_v^2|1 - e^{-2\beta t}|,\end{aligned} \tag{2.5}$$

where we have defined the constant $\rho_v^2 = \gamma_{\|}(\lambda+1)^2/2\beta$. The miraculous cancellation in the nonlinear terms that leads to this strong energy estimate is actually a consequence of the underlying physics; see [9] for details. Since the nonlinear terms uv^* and uw in (2.3)–(2.4) do not influence the change in energy, we say that the nonlinearity is energy preserving. Equation (2.2) admits an equally strong estimate. A smooth solution satisfies

$$(\partial_t + \varepsilon\partial_x)(|u(t,x)|^2) = -2\kappa\varepsilon|u(t,x)|^2 + 2\frac{\varepsilon\kappa}{\mu}\,\mathrm{Re}\,(u^*v)$$
$$\leq -\varepsilon\kappa|u(t,x)|^2 + \frac{\varepsilon\kappa}{\mu^2}|v(t,x)|^2.$$

Integrating this inequality along the characteristic $x - \varepsilon t$ =constant, we have

$$|u(t,x)|^2 \leq e^{-\varepsilon\kappa t}|u(0,x-\varepsilon t)|^2 + \frac{\varepsilon\kappa}{\mu^2}\int_0^t e^{-\varepsilon\kappa(t-s)}|v(s,x-\varepsilon(t-s))|^2 ds. \tag{2.6}$$

Taking the sup over $x \in S^1$, and using the energy estimate (2.5), we obtain

$$\|u(t)\|^2 \leq e^{-\varepsilon\kappa t}\|u(0)\|^2 + \frac{\varepsilon\kappa}{\mu^2}e_{2\beta}(t)\|(v,w)(0)\|^2 + \frac{\rho_v^2}{\mu^2}(1-e^{-\varepsilon\kappa t}), \tag{2.7}$$

where we have defined the exponentially decaying function

$$e_\alpha(t) = \frac{e^{-\varepsilon\kappa t} - e^{-\alpha t}}{\alpha - \varepsilon\kappa}, \tag{2.8}$$

assuming that $\varepsilon\kappa < \alpha$.

These energy estimates will be used to establish the existence of global mild solutions. They also immediately establish the existence of positively invariant regions in $\mathbb{X}$. Trajectories will satisfy $\|(v,w)(t)\| < \|(v,w)(0)\|$, for all $t > 0$, provided

$$\|(v,w)(0)\| > \rho_v. \tag{2.9}$$

Let $c(\varepsilon) = \sup_{t\geq 0}\varepsilon\kappa e_{2\beta}(t)/(1-e^{-\varepsilon\kappa t})$. Since $e_{2\beta}(t) \leq te^{-\varepsilon\kappa t}$ we find that

$$c(\varepsilon) \leq \sup_{y>0}\frac{y}{e^y - 1} = 1.$$

Suppose that the initial conditions satisfy (2.9). The energy estimate (2.7) shows that a sufficient condition for $\|u(t)\| < \|u(0)\|$ for all $t > 0$ is

$$\|u(0)\|^2 > \frac{\|(v,w)(0)\|^2}{\mu^2} + \frac{\rho_v^2}{\mu^2}. \tag{2.10}$$

Conditions (2.9) and (2.10) show that the region

$$\mathcal{D}_0 = \{\|u\|^2 \leq 4\rho_v^2/\mu^2, \|(v,w)\|^2 \leq 2\rho_v^2\} \tag{2.11}$$

is strictly positively invariant. $\mathcal{D}_0$ is also an absorbing region for the flow. The energy estimates (2.5) and (2.7) show that all trajectories enter $\mathcal{D}_0$ at the slow exponential rate $e^{-\varepsilon\kappa t}$ and that the time taken to enter $\mathcal{D}_0$ is uniform on bounded sets.

Remark 2.1. It is important to note that the size of the absorbing region is uniform for $0 < \varepsilon\kappa < 2\beta$. We will use this in our construction of slow manifolds.

Let $(u_i, v_i, w_i), i = 1, 2$, be two smooth solutions. We will estimate the growth of their difference. Define $(\xi,\eta,\zeta) = (u_1,v_1,w_1) - (u_2,v_2,w_2)$ and $(\bar{u},\bar{v},\bar{w}) = ((u_1,v_1,w_1) + (u_2,v_2,w_2))/2$. The differences satisfy

$$\xi_t = \varepsilon\left[-\xi_x + \kappa\left(\frac{\eta}{\mu} - \xi\right)\right], \tag{2.12}$$

$$\eta_t = -(1+i\delta)\eta + \mu(\bar{u}\zeta + \bar{w}\xi), \tag{2.13}$$

$$\zeta_t = -\gamma_{\|}\zeta - \mu\,\mathrm{Re}\,(\bar{u}^*\eta + \bar{v}^*\xi). \tag{2.14}$$

Equations (2.13) and (2.14) give the pointwise error estimate

$$\begin{aligned}\partial_t(|\eta|^2+|\zeta|^2) &= -2|\eta|^2-2\gamma_{\|}|\zeta|^2+2\mu\,\mathrm{Re}\,(\xi\eta^*\bar{w}-\xi\zeta\bar{v}^*)\\ &\le -\beta(|\eta|^2+|\zeta|^2)+\frac{\mu^2|\xi|^2}{\beta}(|\bar{v}|^2+|\bar{w}|^2).\end{aligned}$$

In the second step we have used the elementary inequality $2pq \le \beta p^2 + q^2/\beta$. Integrating and taking the sup over $x \in S^1$, we obtain

$$\|(\eta,\zeta)(t)\|^2 \le e^{-\beta t}\|(\eta,\zeta)(0)\|^2+\frac{\mu^2}{\beta}\int_0^t e^{-\beta(t-s)}\|\xi(s)\|^2\|(\bar{v},\bar{w})(s)\|^2 ds. \tag{2.15}$$

The definition of $(\bar{v},\bar{w})$, combined with the energy estimate (2.5), gives

$$\|(\bar{v},\bar{w})(t)\|^2 \le C \quad \forall t\ge 0, \tag{2.16}$$

where C is a constant that is uniform for initial conditions in any fixed ball.

An energy estimate for ξ can be obtained from (2.12). For two smooth solutions we have

$$\begin{aligned}(\partial_t+\varepsilon\partial_x)(|\xi|^2) &= 2\varepsilon\kappa\left(-|\xi|^2+\frac{1}{\mu}\,\mathrm{Re}\,(\xi^*\eta)\right)\\ &\le 2\varepsilon\kappa\left(-|\xi|^2+\frac{|\xi|^2}{2}+\frac{|\eta|^2}{2\mu^2}\right) = -\varepsilon\kappa|\xi|^2+\frac{\varepsilon\kappa}{\mu^2}|\eta|^2.\end{aligned}$$

Integrating this inequality along the characteristic $x-\varepsilon t$ =constant, we have

$$|\xi(t,x)|^2 \le e^{-\varepsilon\kappa t}|\xi(0,x-\varepsilon t)|^2+\frac{\varepsilon\kappa}{\mu^2}\int_0^t e^{-\varepsilon\kappa(t-s)}|\eta(s,x-\varepsilon(t-s))|^2 ds,$$

and taking the sup over $x\in S^1$ we obtain

$$\|\xi(t)\|^2 \le e^{-\varepsilon\kappa t}\|\xi(0)\|^2+\frac{\varepsilon\kappa}{\mu^2}\int_0^t e^{-\varepsilon\kappa(t-s)}\|(\eta,\zeta)(s)\|^2 ds. \tag{2.17}$$

Since the expressions $\|\xi(t)\|^2$ and $\|(\eta(t),\zeta(t))\|^2$ occur often below, we now introduce separate notation for them. Let

$$a(t)=\|\xi(t)\|^2, \quad b(t)=\|(\eta,\zeta)(t)\|^2. \tag{2.18}$$

Combining the inequalities (2.15), (2.16), and (2.17) with the notation of (2.18), we obtain

$$\begin{aligned}b(t) &\le b(0)e^{-\beta t}+\frac{C\mu^2}{\beta}\int_0^t e^{-\beta(t-s)}\left(e^{-\varepsilon\kappa s}a(0)+\frac{\varepsilon\kappa}{\mu^2}\int_0^s e^{-\varepsilon\kappa(s-\tau)}b(\tau)d\tau\right)ds\\ &= b(0)e^{-\beta t}+\frac{C\mu^2}{\beta}a(0)e_\beta(t)+\frac{C\varepsilon\kappa}{\beta}\int_0^t\int_0^s e^{-\beta(t-s)-\varepsilon\kappa(s-\tau)}b(\tau)d\tau ds\\ &= b(0)e^{-\beta t}+\frac{C\mu^2}{\beta}a(0)e_\beta(t)+\frac{C\varepsilon\kappa}{\beta}\int_0^t b(\tau)e^{-\beta t+\varepsilon\kappa\tau}\int_\tau^t e^{(\beta-\varepsilon\kappa)s}ds d\tau.\end{aligned}$$

Computing the inner integral we obtain the estimate

$$b(t)\le b(0)e^{-\beta t}+\frac{C\mu^2}{\beta}a(0)e_\beta(t)+\frac{C\varepsilon\kappa}{\beta}\int_0^t e_\beta(t-\tau)b(\tau)d\tau. \tag{2.19}$$

Here $e_\beta(t)$ is defined as in (2.8). Thus, we have $e_\beta(t) \le te^{-\varepsilon\kappa t}$ for positive times (we suppose that $0 < \varepsilon\kappa < \beta$). As in [9], we apply Gronwall's inequality to (2.19), and use the resulting estimate in (2.17) to obtain

$$\sup_{t\in[0,T]} (a(t) + b(t)) \le C(T, \|(u_i, v_i, w_i)(0)\|)(a(0) + b(0)).$$

We have made the assumption that $t \ge 0$ for simplicity. One may work through the estimates again to find that for any fixed $T > 0$,

$$\sup_{t\in[-T,T]} (a(t) + b(t)) \le C(T, \|(u_i, v_i, w_i)(0)\|)(a(0) + b(0)). \tag{2.20}$$

2.3. Existence of a smooth flow. We now define precisely the dynamical system we will be studying and then prove the existence of a smooth global flow.

Definition 2.2. *$(u(t), v(t), w(t)) \in \mathbb{X}$ is a* mild solution *to the laser equations* (2.2)–(2.4) *if it satisfies the integral equations*

$$u(t) = e^{-\varepsilon\kappa t}e^{-\varepsilon t\partial_x}u(0) + \frac{\varepsilon\kappa}{\mu}\int_0^t e^{-\varepsilon\kappa(t-s)}e^{-\varepsilon(t-s)\partial_x}v(s)ds, \tag{2.21}$$

$$v(t) = e^{-(1+i\delta)t}v(0) + \mu\int_0^t e^{-(1+i\delta)(t-s)}u(s)w(s)ds, \tag{2.22}$$

$$w(t) = e^{-\gamma_\| t}w(0) + (\lambda + 1)(1 - e^{-\gamma_\| t}) - \mu\int_0^t e^{-\gamma_\|(t-s)}\,\mathrm{Re}\,(u(s)^*v(s))ds. \tag{2.23}$$

The notation $e^{-\varepsilon t\partial_x}$, with $t \in \mathbb{R}$, refers to the one parameter linear group generated by the wave equation $u_t + \varepsilon u_x = 0$ in $C(S^1;\mathbb{C})$. It is defined by the shift map $(e^{-\varepsilon t\partial_x}u)(x) = u(x - \varepsilon t)$.

Remark 2.3. The integrals in (2.21)–(2.23) are interpreted as elements in $\mathbb{X}$. Since we are considering continuous functions, the integrals are well defined if and only if they are defined at each point $x \in S^1$. Notice that the product $u(s)w(s)$ is a well-defined continuous function. In [9] the laser equations do not admit a variation of constants formula in L^2. In our work a variation of constants formula is essential.

Remark 2.4. The integral equations (2.22) and (2.23) are equivalent to the differential equations (2.3) and (2.4) since the right-hand side of the differential equations contains no unbounded operator. Thus the a priori estimates (2.5) and (2.15) apply to *all* mild solutions, not just smooth solutions. The a priori estimates on u and ξ, (2.7) and (2.17), are extended to all mild solutions by approximating continuous functions with C^1 functions.

Theorem 2.5 (C^∞ flow). *The laser equations define a C^∞ global flow in the sense of mild solutions. That is, there exists a C^∞ map $\Phi : \mathbb{R} \times \mathbb{X} \to \mathbb{X}$ with the following properties:*

(a) *$\Phi(t, u_0, v_0, w_0)$ is the unique solution to* (2.21)–(2.23) *with initial conditions $\Phi(0, u_0, v_0, w_0) = (u_0, v_0, w_0)$.*

(b) *The set of maps $\varphi_t : \mathbb{X} \to \mathbb{X}, t \in \mathbb{R}$ defined by $\varphi_t(u_0, v_0, w_0) = \Phi(t, u_0, v_0, w_0)$ is a one parameter group of C^∞ diffeomorphisms of $\mathbb{X}$.*

Sketch of the proof. A contraction mapping argument shows that for every point in $\mathbb{X}$ within the ball of radius ρ there is a unique mild solution defined for a time interval $[-T(\rho), T(\rho)]$. A well-known theorem of Segal [26] asserts that solutions fail to exist after a finite time, T_{crit}, if and only if they blow up, i.e., $\|(u(t), v(t), w(t))\| \to \infty$ as $t \to T_{crit}$. The a priori estimates (2.5) and (2.7) show that this is impossible.

Thus through every point (u_0, v_0, w_0) there is a unique solution for all $t \in \mathbb{R}$ denoted by $\Phi(t, u_0, v_0, w_0)$ with $\Phi(0, u_0, v_0, w_0) = (u_0, v_0, w_0)$. Let φ_t be defined as in (b). Clearly, $\varphi_0 = \mathrm{Id}$. Then (2.20) shows that each φ_t is continuous (in fact, locally Lipschitz). The group property follows from uniqueness of solutions.

The proof that the flow is C^∞ is by induction on the order of the derivative. Each step of the argument follows. For any positive integer r, formal differentiation of the equations for the $(r-1)$th derivative yields a linear integral equation that the rth derivative must satisfy (for $r = 1$, we differentiate (2.21)–(2.23)). The existence of a unique solution to this integral equation on a time interval $[-T(\rho, r), T(\rho, r)]$ is proven by a contraction mapping argument. Gronwall estimates show that the derivative grows at worst exponentially in time. Thus, the derivatives of the flow are defined for all $t \in \mathbb{R}$. This is a standard calculation (see, e.g., [6, 18]) and we omit the details. The heart of the matter is that the nonlinear terms on the right-hand side of (2.22)–(2.23) are smooth, and thus all derivatives exist. □

Remark 2.6. We did not need the full strength of the estimates for differences (2.15)–(2.17) in this proof. The estimates will be used in section 4 to prove the cone property.

2.4. Asymptotic dynamics.

The laser equations are dissipative. All trajectories must enter the trapping region $\mathcal{D}_0$ in finite time. To capture the asymptotic behavior of the system, we define the global attractor

$$\mathcal{A} = \bigcap_{t \geq 0} \varphi_t(\mathcal{D}_0).$$

Since $\mathcal{D}_0$ is absorbing and closed, this agrees with the definition of the attractor as the ω-limit set of the absorbing ball

$$\omega(\mathcal{D}_0) = \bigcap_{T \geq 0} \overline{\bigcup_{t \geq T} \varphi_t(\mathcal{D}_0)}.$$

Although the flow is dissipative, it is not smoothing, and it is not obvious that this definition of the attractor is meaningful. However, this follows from the asymptotic smoothing property of the laser equations proved by Constantin, Foias, and Gibbon [9]; see also [23]. Let $\mathcal{B}$ denote the attractor in L^2. The main result of Constantin, Foias, and Gibbon is that $\mathcal{B}$ is composed of C^∞ functions and that it has finite Hausdorff dimension. Thus it also has finite topological dimension. In [23] the regularity result was improved: The attractor $\mathcal{B}$ is in every Gevrey class G^s for $s > 1$, i.e., the attractor is "almost analytic." Furthermore, the estimates in [9, 23] show that the attractor is compact by the Arzela–Ascoli theorem. Since the inclusion $\iota : \mathbb{X} \to L^2$ is continuous, these results apply immediately to the flow in $\mathbb{X}$. Applying the regularity result we see that $\iota(\mathcal{A}) = \mathcal{B}$. Furthermore, since $\mathcal{B}$ is compact, the inverse map restricted to $\mathcal{B}$ is continuous. Hence $\mathcal{A}$ and $\mathcal{B}$ are homeomorphic and have the same topological dimension.

These theorems are independent of the scaling assumptions of our paper. We assert that under suitable scaling hypotheses, one can simplify the geometry further by constructing a normally hyperbolic invariant manifold that contains the attractor.

3. Geometry in the limit $\varepsilon = 0$.

If $\varepsilon = 0$, then $u_t = 0$ in (2.2). By inspection one sees the existence of a manifold of equilibria, $\mathcal{M}_0$, given as the graph of a map $h : \mathbb{X}_1 \to \mathbb{X}_2$. We denote its components by $h(u) = (h_v(u), h_w(u))$. These maps are defined pointwise for $x \in S^1$ by

$$h_v(u)(x) = \mu(1-i\delta)\frac{(\lambda+1)u(x)}{1+\delta^2+\lambda|u(x)|^2}, \quad h_w(u)(x) = \frac{(1+\delta^2)(\lambda+1)}{1+\delta^2+\lambda|u(x)|^2}. \tag{3.1}$$

For large $|u(x)|$ the denominator dominates; therefore, $\mathcal{M}_0$ is uniformly bounded. The pointwise maps $h_v(u)(x)$ and $h_w(u)(x)$ are C^∞ as functions of $u(x)$. Since pointwise operations extend naturally in $C(S^1)$, we find that h is C^∞ as a map between $\mathbb{X}_1 \to \mathbb{X}_2$. Thus, $\mathcal{M}_0$ is a C^∞ manifold.

In this limit we can solve the laser equations explicitly. We split (u, v) into their real and imaginary parts, i.e., $(u, v) = (\mathrm{Re}(u), \mathrm{Re}(v)) + i(\mathrm{Im}(u), \mathrm{Im}(v))$, and then rewrite (2.3)–(2.4) as

$$\partial_t \begin{pmatrix} \mathrm{Re}(v) \\ \mathrm{Im}(v) \\ w \end{pmatrix} = A(u) \begin{pmatrix} \mathrm{Re}(v) \\ \mathrm{Im}(v) \\ w \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \gamma_\| (\lambda+1) \end{pmatrix}, \tag{3.2}$$

where $A(u)$ is the bounded multiplication operator defined by

$$A(u) = \begin{pmatrix} -1 & \delta & \mu\mathrm{Re}(u) \\ -\delta & -1 & \mu\mathrm{Im}(u) \\ -\mu\mathrm{Re}(u) & -\mu\mathrm{Im}(u) & -\gamma_\| \end{pmatrix}. \tag{3.3}$$

Thus, the solution to the laser equations (2.3)–(2.4) in this limit is $u = u(x)$, and

$$\begin{pmatrix} \mathrm{Re}(v)(t) \\ \mathrm{Im}(v)(t) \\ w(t) \end{pmatrix} = e^{tA(u)} \begin{pmatrix} \mathrm{Re}(v)(0) \\ \mathrm{Im}(v)(0) \\ w(0) \end{pmatrix} + \int_0^t e^{(t-s)A(u)} \begin{pmatrix} 0 \\ 0 \\ \gamma_\| (\lambda+1) \end{pmatrix} ds. \tag{3.4}$$

Here u is treated as a parameter and the fibers of constant u are invariant under the flow. Within each fiber, trajectories decay to the equilibrium $(u, h(u))$. The next lemma states that the decay rate is uniform over $\mathcal{M}_0$.

LEMMA 3.1. $\|e^{tA(u)}\| \le e^{-\beta t}$ *for all* $u \in \mathbb{X}_1$.

Proof. This follows from an estimate similar to (2.5). The operator $A(u)$ is broken into two parts: a diagonal matrix that is independent of u and a skew matrix that depends on u. The skew matrix does not influence the growth or decay of energy, and hence u cannot influence the decay in $\|(v, w)(t)\|$. □

Clearly, Lemma 3.1 reflects a strong stability of $\mathcal{M}_0$ that depends on the skew nonlinearity. As we have emphasized earlier, this is actually a consequence of the underlying physics. Figure 3.1 describes the geometry of the flow with two key geometric objects. The first is the critical manifold $\mathcal{M}_0$, the second is the smooth invariant family $\mathcal{F}_{u_0} := \{(u, v, w) | u = u_0\}$ parametrized by $u_0 \in \mathbb{X}_1$. There is a purely metric characterization of $\mathcal{F}_{u_0}$: For any $0 < \gamma < \beta$ these manifolds are γ-stable manifolds in the sense of Chow, Lin, and Lu [7]; i.e., for $t \in \mathbf{R}_+$ and fixed $(u_0, h(u_0))$ the set of points $\{(u, v, w) : \|\varphi_t(u, v, w) - \varphi_t(u_0, h(u_0))\| = O(e^{-\gamma t})\}$ is identical to $\mathcal{F}_{u_0}$. For $\varepsilon > 0$ the system is dissipative in the $\mathbb{X}_1$ direction as well, and all trajectories are sucked into the absorbing region $\mathcal{D}_0$. Thus, it is sufficient to show that $\mathcal{M}_0$ and $\mathcal{F}_{u_0}$ persist within $\mathcal{D}_0$. Roughly speaking, we shall show that there is an $\varepsilon_* > 0$ so that for all $0 \le \varepsilon \le \varepsilon_*$, there is a smooth (but not C^∞) invariant manifold $\mathcal{M}_\varepsilon$ given as a graph $(u, h_\varepsilon(u))$ over $\Pi_1(\mathcal{D}_0)$ that contains the asymptotic dynamics (in particular the attractor $\mathcal{A}$) and is exponentially attracting.

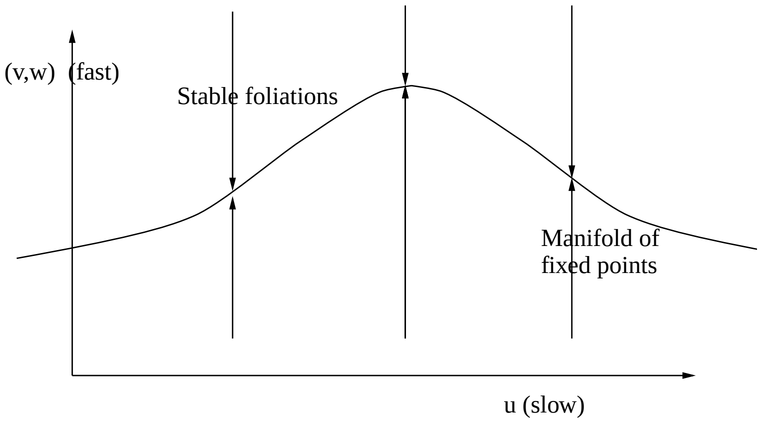


FIG. 3.1. *Geometry in the singular limit $\varepsilon = 0$.*

4. Existence of the invariant manifold.

4.1. The main theorem.

THEOREM 4.1. *For any integer r, there is an $\varepsilon_*(r) > 0$ so that for each $\varepsilon \in [0, \varepsilon_*(r)]$ there is a positively invariant C^r manifold, $\mathcal{M}_\varepsilon$, given as a graph over $\Pi_1(\mathcal{D}_0)$. This manifold attracts all initial conditions exponentially fast and contains the attractor $\mathcal{A}$ of the Maxwell–Bloch equations.*

Sections 4 and 5 are devoted to a proof of the main theorem. The consequences of this theorem are explored in section 6.

4.2. The modified equations. We will use Hadamard's graph transform method to prove the existence of a persisting manifold [14]. We will restrict our attention to the flow within an absorbing ball and modify the nonlinearity outside this ball. This approach has been used to prove the existence of finite-dimensional attracting manifolds for dissipative dynamical systems (e.g., reaction diffusion equations) [13].

Let $R_1 = 2\rho_v/\mu$ and $R_2 = \sqrt{2}\rho_v$. Then R_1 and R_2 are sufficiently large that the region

$$\mathcal{D} = \{\|u\|_{\mathbb{X}_1} \leq 2R_1, \|(v, w)\|_{\mathbb{X}_2} \leq 2R_2\} \tag{4.1}$$

is absorbing and positively invariant (see (2.11) and the discussion preceding it). We denote this region by $\mathcal{D}$ and note that $\mathcal{D} = 2\mathcal{D}_0$.

Let $\chi_1 : \mathbb{C} \to [0, 1]$ be a C^∞ function with compact support that takes the values $\chi_1(s) = 1,\ 0 \leq |s| \leq 1,\ \chi_1(s) = 0$ for $2 \leq |s| < \infty$ and has uniformly bounded derivative $|D\chi_1(s)| \leq 2$. Let $\chi_2 : \mathbb{C} \times \mathbb{R} \to [0, 1]$ be a cut-off function with analogous properties. Define the cut-off functions $\chi_{R_i} : \mathbb{X}_i \to [0, 1]$ by $\chi_{R_1}(u)(x) = \chi_1(u(x)/R_1)$ and $\chi_{R_2}(v, w)(x) = \chi_2((v(x), w(x))/R_2)$. One may prove that χ_{R_i}, $i = 1, 2$, are C^∞. As is common in invariant manifold theory, we will modify the laser equations so as to obtain global estimates. Let

$$\kappa\left(\frac{v}{\mu} - u\right)\chi_{R_1}(u)\chi_{R_2}(v, w) = g(u, v, w) \quad \text{and} \quad u\chi_{R_1}(u) = f(u). \tag{4.2}$$

Consider the modified laser equations

$$u_t = -\varepsilon u_x + \varepsilon g(u, v, w), \tag{4.3}$$

$$v_t = -(1 + i\delta)v + \mu f(u)w, \tag{4.4}$$

$$w_t = \gamma_\| (\lambda + 1 - w) - \mu \mathrm{Re}(f(u)v^*). \tag{4.5}$$

We modify only the u term in the nonlinearity in (2.3)–(2.4). This allows us to retain an estimate similar to the energy estimate (2.15).

LEMMA 4.1. *For* $(u_i, v_i, w_i) \in \mathbb{X}$, $i = 1, 2$, *we have*

(a) $\|f(u_1, v_1, w_1) - f(u_2, v_2, w_2)\| \le 5\|u_1 - u_2\|$,

(b) $\|g(u_1, v_1, w_1) - g(u_2, v_2, w_2)\| \le \kappa(5+4\sqrt{2})(\|u_1 - u_2\| + \mu^{-1}\|(v_1, w_1) - (v_2, w_2)\|)$.

Proof. Without loss of generality suppose $\max(|u_i(x)|) = |u_2(x)|$. If $|u_2(x)| \le 2R_1$,

$$\begin{aligned} &|u_1\chi_{R_1}(u_1)(x) - u_2\chi_{R_1}(u_2)(x)| \\ &\le |u_1(x) - u_2(x)||\chi_{R_1}(u_1)(x)| + |u_2(x)||\chi_{R_1}(u_1)(x) - \chi_{R_1}(u_2)(x)| \\ &\le |u_1(x) - u_2(x)| + 2R_1 \frac{2}{R_1}|u_1(x) - u_2(x)| \le 5\|u_1 - u_2\|. \end{aligned}$$

If $\min(|u_1(x)|, |u_2(x)|) > 2R_1$, the above inequality is trivial since the left-hand side is zero. Finally, if $|u_2(x)| > 2R_1$ and $|u_1(x)| \le 2R_1$, we have

$$|u_1\chi_{R_1}(u_1)(x) - u_2\chi_{R_1}(u_2)(x)| = |u_1\chi_{R_1}(u_1)(x) - u_1\chi_{R_1}(u_2)(x)| \le 4\|u_1 - u_2\|. \tag{4.6}$$

Taking the sup over x we obtain (a). Similar calculations show that the difference in g is bounded by

$$\kappa\left(\left(5 + \frac{4R_2}{\mu R_1}\right)\|u_1 - u_2\| + \left(\frac{5}{\mu} + \frac{4R_1}{R_2}\right)\|(v_1, w_1) - (v_2, w_2)\|\right).$$

But $R_2/R_1 = \mu/\sqrt{2}$. Simplifying the above estimate, we obtain (b). □

Remark 4.2. We make the following important observation regarding the modified flow. Suppose $\|u(0)\| > 2R_1$. Then there exists an open interval I in S^1 so that $|u(0)(x)| > 2R_1$ for all $x \in I$, and hence $g((u, v, w)(0)) = 0$, on this interval. Integrating (4.3) along the characteristic $x - \varepsilon t =$ constant, we find that $u(t, x)$ is constant on the characteristics through $I \times \{t = 0\}$. Thus, $\|u(t)\| > 2R_1$ for all $t \in \mathbb{R}$, and the region $\{\|u\| > 2R_1\}$ in phase space is invariant for the modified flow. This implies its complement is also invariant. Hence the phase space splits into two *invariant* regions, the closed cylinder $\{\|u\| \le 2R_1\}$ and its exterior.

Remark 4.3. Within the region $\{\|u\| \le R_1, \|(v, w)\| \le R_2\}$ the modified and unmodified equations agree on a dense set, and hence their flows agree locally in time. But by the choice of R_i, this region is positively invariant, and thus the flows agree for all positive time. As a result they have identical asymptotic dynamics within this region. We will prove the following invariant manifold theorem for the mild formulation of the modified equations (4.3)–(4.5). The mild formulation is

$$u(t) = e^{-\varepsilon t\partial_x}u(0) + \varepsilon\int_0^t e^{-\varepsilon(t-s)\partial_x} g(u(s), v(s), w(s))ds, \tag{4.7}$$

$$v(t) = e^{-(1+i\delta)t}v(0) + \mu\int_0^t e^{-(1+i\delta)(t-s)} f(u(s))w(s)ds, \tag{4.8}$$

$$\begin{aligned} w(t) = {} & e^{-\gamma_\| t}w(0) + (\lambda + 1)(1 - e^{-\gamma_\| t}) \\ & -\mu\int_0^t e^{-(t-s)\gamma_\|} \mathrm{Re}\,(f(u(s))^* v(s))ds. \end{aligned} \tag{4.9}$$

Theorem 4.4. *For any integer r, there exists an $\varepsilon_*(r) > 0$ so that for each $\varepsilon \in [0, \varepsilon_*(r)]$ there is a C^r manifold, $\mathcal{M}_\varepsilon$, invariant under the flow of the modified Maxwell–Bloch equations* (4.7)–(4.9). *$\mathcal{M}_\varepsilon$ is given as a graph over $\Pi_1(\mathcal{D})$. This manifold attracts all points in the absorbing region exponentially fast and contains the attractor $\mathcal{A}$ of the Maxwell–Bloch equations* (2.21)–(2.23).

Theorem 4.4 implies Theorem 4.1 because, by Remark 4.3, the asymptotic dynamics of modified and unmodified systems agree within $\mathcal{D}_0$. Since $\mathcal{D}_0$ is only positively invariant, the invariance of the manifold in Theorem 4.4 is weakened to positive invariance in Theorem 4.1.

4.3. A priori estimates. We reconsider the a priori estimates of section 2 in light of the above modifications. Henceforth, in sections 4 and 5, φ_t denotes the flow of the modified equations (4.7)–(4.9). In all that follows, we will only consider trajectories that start within the positively invariant region $\mathcal{D}$. Thus the constants, C_j, that occur in inequalities will generally depend on R_i and the parameters $\kappa, \lambda, \gamma_\|$, and μ. We also assume that the time t is positive.

Remark 4.2 implies the uniform bound

$$\|u(t)\| \leq 2R_1, \quad t \in \mathbb{R}, \tag{4.10}$$

for all trajectories starting within $\mathcal{D}$. The modification has also been chosen so that the energy estimate (2.15) is unchanged (i.e., we retain the cancellation of nonlinear terms). Thus, by the choice of R_2 trajectories starting within $\mathcal{D}$ satisfy the uniform bound

$$\|(v, w)(t)\| \leq 2R_2, \quad t \geq 0. \tag{4.11}$$

Estimates for differences between trajectories are derived as in section 2. As in (2.17) we have

$$a(t) \leq e^{C_1 \varepsilon t} a_0 + C_2 \varepsilon \int_0^t e^{C_1 \varepsilon (t-s)} b(s) ds \tag{4.12}$$

for $C_i = C_i(\kappa, \mu, R_i), i = 1, 2$. The analogue of (2.15) is derived from (4.3) and (4.4). The differences (η, ζ) now satisfy

$$\begin{aligned} \partial_t(|\eta|^2 + |\zeta|^2) &= -2|\eta|^2 - 2\gamma_\| |\zeta|^2 + 2\mathrm{Re}\,(\eta^*(f(u_1)w_1 - f(u_2)w_2)) \\ &\quad -2\mathrm{Re}\,(\zeta(f(u_1)v_1^* - f(u_2)v_2^*)) \\ &= -2|\eta|^2 - 2\gamma_\| |\zeta|^2 + \mathrm{Re}\,((f(u_1) - f(u_2))(\bar{w}\eta^* - \bar{v}^*\zeta)). \end{aligned} \tag{4.13}$$

Notice that the choice of the modification is such that the term involving $f(u_1)+f(u_2)$ cancels. This is important as it ensures that we retain the uniform decay normal to the manifold $\mathcal{M}_0$, independent of the basepoint u. One can now use Lemma 4.1 and the energy estimate (4.11) in (4.13) and integrate to find

$$b(t) \leq e^{-\beta t} b(0) + C_3 \int_0^t e^{-\beta(t-s)} a(s) ds. \tag{4.14}$$

The constant C_3 depends only on the parameters $\kappa, \mu, \lambda, \beta$ and the radii R_i. We also need lower estimates on $a(t)$ and $b(t)$ that are derived similarly. For example, (4.3) yields

$$(\partial_t + \varepsilon \partial_x)|\xi(t, x)|^2 \geq -C\varepsilon|\xi|(|\xi| + |\eta|) \geq -C_1 \varepsilon |\xi|^2 - C_2 \varepsilon |\eta|^2,$$

so that integrating between $t_1 \leq t_2$ and taking the sup over x we have

$$a(t_2) \geq e^{-C_1\varepsilon(t_2-t_1)}a(t_1) - C_2\varepsilon \int_{t_1}^{t_2} e^{-C_1\varepsilon(t_2-s)}b(s)ds. \tag{4.15}$$

Finally, from (4.13) and the energy estimate (4.11) we have the pointwise inequality

$$\partial_t(|\eta(t,x)|^2 + |\zeta(t,x)|^2) \geq -3\tilde{\beta}(|\eta|^2 + |\zeta|^2) - C|\xi|^2, \tag{4.16}$$

where we have defined $\tilde{\beta} = \max(1, \gamma_{\|})$. Integrating this inequality and taking the sup over $x \in S^1$ we obtain

$$b(t_2) \geq e^{-3\tilde{\beta}(t_2-t_1)}b(t_1) - C_4 \int_{t_1}^{t_2} e^{-3\tilde{\beta}(t_2-s)}a(s)ds. \tag{4.17}$$

These a priori estimates can be used to prove the existence of a C^∞ flow for the dynamical system defined by (4.7)–(4.9) as in Theorem 2.5. We will not state a separate theorem.

4.4. The cone property. The graph transform will be defined by applying the map φ_T to Lipschitz sections of the normal bundle of the critical manifold $\mathcal{M}_0$. Over sufficiently large time we expect the flow to contract strongly in the normal direction. This is made precise in the cone condition formulated by Conley, and used since then by several authors. It is an essential geometric feature in the persistence theorem of Bates, Lu, and Zeng and a comprehensive list of references may be found in their article [4].

Choose $T > 0$ so that

$$e^{-\beta T/2} = \frac{1}{32}. \tag{4.18}$$

T will be held fixed in all that follows. In the following propositions ε_* denotes an upper limit that may only decrease from one assertion to the next. This follows the convention in [4]. For $(u, v, w) \in \mathcal{D}$, we will use the cone

$$K_L(u,v,w) = \{(u_1,v_1,w_1) \in \mathcal{D} : \|(v_1,w_1) - (v,w)\|_{\mathbb{X}_2} \leq L\|u_1 - u\|_{\mathbb{X}_1}\}. \tag{4.19}$$

Lemma 4.2 (the moving cone lemma). *There exists $\varepsilon_* > 0$ and $L > 0$ such that for $\varepsilon \in [0, \varepsilon_*]$, $t \in [0,T]$, and each point $(u,v,w) \in \mathcal{D}$, the cone $K_L(u,v,w)$ is carried by the diffeomorphism φ_t into the cone $K_L(\varphi_t(u,v,w))$.*

Remark 4.5. The statement of Lemma 4.2 is uniform over all points in the absorbing region. Geometrically, this implies a squeezing property of the flow.

Proof. $\mathcal{D}$ is positively invariant: thus for any $(u,v,w) \in \mathcal{D}$, $L > 0$, and $t \geq 0$, φ_t carries the cone $K_L(u,v,w)$ into $\mathcal{D}$. It remains to prove that for suitable $L > 0$, if two trajectories start in $\mathcal{D}$ and satisfy $b_0 \leq L^2 a_0$, then $b(t) \leq L^2 a(t)$ for all $t \in [0,T]$. Since the initial conditions lie in $\mathcal{D}$, $a(t)$ and $b(t)$ must satisfy the a priori estimates (4.12) and (4.14). Our proof will demonstrate a technique of dealing with these coupled inequalities by exploiting the gap in the exponential rates.

For any $\gamma \in (C_1\varepsilon, \beta)$ we define $|a|_{\gamma,t} = \sup_{s\in[0,t]} a(s)e^{\gamma s}$. Similarly, we define $|b|_{\gamma,t}$. It follows that $|a|_{\gamma,t}$ is an increasing function of t. We will use $\gamma = \beta/2$, though the argument will work for any γ that satisfies the gap condition $C_1\varepsilon < \gamma < \beta$. We further assume that ε_* is so small that $C_1\varepsilon_* < \beta/2$.

We multiply (4.12) by $e^{\beta s/2}$ to obtain

$$\begin{aligned} a(s)e^{\beta s/2} &\leq e^{(\beta/2+C_1\varepsilon)s}a_0 + C_2\varepsilon\int_0^s e^{(\beta/2+C_1\varepsilon)(s-\tau)}e^{\beta\tau/2}b(\tau)d\tau \\ &\leq e^{(\beta/2+C_1\varepsilon)s}a_0 + \frac{C_2\varepsilon}{\beta/2+C_1\varepsilon}\left(e^{(\beta/2+C_1\varepsilon)s}-1\right)|b|_{\beta/2,s} \\ &\leq e^{(\beta/2+C_1\varepsilon)s}\left(a_0+\varepsilon C_2 s|b|_{\beta/2,s}\right). \end{aligned}$$

In the last step we have used the elementary inequality $1-e^{-t}\leq t$ for positive t. Taking the sup over $s\in[0,t]$, and using the fact that $|a|_{\gamma,s}$ is an increasing function of s, we obtain

$$|a|_{\beta/2,t}\leq e^{(\beta/2+C_1\varepsilon)t}\left(a_0+C_2\varepsilon t|b|_{\beta/2,t}\right). \tag{4.20}$$

We apply a similar calculation to (4.14) to obtain

$$\begin{aligned} b(s)e^{\beta s/2} &\leq b_0e^{-\beta s/2}+C_3\int_0^s e^{-\beta(s-\tau)/2}e^{\beta\tau/2}a(\tau)d\tau \\ &\leq b_0+C_3\frac{(1-e^{-\beta s/2})}{\beta/2}|a|_{\beta/2,s}\leq b_0+C_3s|a|_{\beta/2,s}. \end{aligned}$$

Taking the sup over $s\in[0,t]$ we find

$$|b|_{\beta/2,t}\leq b_0+C_3t|a|_{\beta/2,t}. \tag{4.21}$$

Combining the inequalities (4.20) and (4.21) we find

$$|b|_{\beta/2,t}\leq b_0+C_3te^{(\beta/2+C_1\varepsilon)t}\left(a_0+C_2\varepsilon t|b|_{\beta/2,t}\right). \tag{4.22}$$

We suppose that ε_* is chosen so small that for all $\varepsilon\in[0,\varepsilon_*]$, we have

$$\varepsilon e^{(\beta/2+C_1\varepsilon)t}C_2C_3t^2\leq\frac{1}{2}. \tag{4.23}$$

Then using the hypothesis $b_0\leq l^2a_0$, and (4.23) in (4.22) we find

$$|b|_{\beta/2,t}\leq a_0\left[\frac{l^2+C_3te^{(\beta/2+C_1\varepsilon)t}}{1-\varepsilon C_2C_3t^2e^{(\beta/2+C_1\varepsilon)t}}\right]=:a_0\theta(t,\varepsilon), \tag{4.24}$$

where we have defined a new function $\theta(t,\varepsilon)$ to simplify notation. Furthermore, we set $t_1=0$, and $t_2=t$ in the backward time estimate (4.15) to deduce that

$$\begin{aligned} a_0 &\leq e^{C_1\varepsilon t}a(t)+C_2\varepsilon\int_0^t e^{C_1\varepsilon s}b(s)ds \\ &\leq e^{C_1\varepsilon t}\left(a(t)+C_2\varepsilon t|b|_{\beta/2,t}\right). \end{aligned} \tag{4.25}$$

Thus, combining (4.24) and (4.25), we have

$$|b|_{\beta/2,t}\leq\theta(t,\varepsilon)a_0\leq\theta(t,\varepsilon)e^{C_1\varepsilon t}\left(a(t)+C_2\varepsilon t|b|_{\beta/2,t}\right). \tag{4.26}$$

We reduce ε_* if necessary so that $\sup_{t\in[0,T]}\varepsilon C_2t\theta(t,\varepsilon)e^{C_1\varepsilon t}\leq 1/2$. Then we have

$$b(t)\leq e^{-\beta/2t}|b|_{\beta/2,t}\leq\frac{\theta(t,\varepsilon)e^{-(\beta/2-C_1\varepsilon)t}}{1-C_2\varepsilon t\theta(t,\varepsilon)e^{C_1\varepsilon t}}a(t). \tag{4.27}$$

Thus, the cone condition (i.e., $b(t) \leq L^2 a(t)$) will be satisfied if we ensure that for all $t \in [0,T]$ we have

$$\tilde{\theta}(t,\varepsilon) := \frac{\theta(t,\varepsilon)e^{-(\beta/2 - C_1\varepsilon)t}}{1 - C_2\varepsilon t\theta(t,\varepsilon)e^{C_1\varepsilon t}} - L^2 \leq 0. \tag{4.28}$$

The function $\tilde{\theta}(t,\varepsilon)$ is smooth in t and ε for $0 \leq t \leq T, 0 \leq \varepsilon \leq \varepsilon_*$, since by the choice of ε_* the denominator is bounded away from zero. Notice that if we let $t = 0$ in (4.24) we have $\theta(0,\varepsilon) = L^2$; hence $\tilde{\theta}(0,\varepsilon) = 0$. If $\varepsilon = 0$, then the inequality (4.28) reduces to

$$\tilde{\theta}(t,0) = -L^2(1 - e^{-\beta t/2}) + C_3 t \leq 0. \tag{4.29}$$

Thus we choose

$$L^2 \geq 2C_3 \max\left(2/\beta, T(1 - e^{-\beta T/2})^{-1}\right) = 2C_3 T\frac{32}{31} \tag{4.30}$$

(see 4.18). This choice ensures that $\tilde{\theta}(t,0)$ is a decreasing function of t in the range $[0,T]$ and the inequality (4.29) is an equality only at $t = 0$. But then to show that (4.28) is true for small positive ε, it suffices to ascertain its validity near $t = 0$. The choice of L in (4.30) ensures that the slope

$$\frac{d\tilde{\theta}(t,0)}{dt}|_{t=0} \leq -C_3 < 0,$$

which implies that for sufficiently small ε_* the inequality $\max_{t\in[0,T]} \tilde{\theta}(t) \leq 0$ is satisfied. In other words, $b(t) \leq L^2 a(t)$ for all $t \in [0,T]$. □

Remark 4.6. To simplify some estimates later, we further suppose that

$$L^2 = 8\max(C_3 T, C_9), \tag{4.31}$$

where C_9 is a constant that occurs in the proof of Lemma 5.1. This simplifies some estimates in the proof of existence and smoothness of the slow manifold $\mathcal{M}_\varepsilon$.

A point about the proof that an expert may find strange is the use of direct estimates on the flow as opposed to estimates from the linearization near the manifold. The laser equations admit strong estimates which is why this approach works. Typically, the best one can do is obtain a cone condition in a neighborhood of the manifold. Another unusual feature is the use of a Lipschitz constant L that is not small. In Fenichel's work [11] the slope of the Lipschitz sections (i.e., L) is small. The distinction is that we use a single coordinate chart for $\mathcal{M}_0$, so L is finite to account for the nonzero slope of $\mathcal{M}_0$. This is to avoid a global coordinate transformation that would lead to vexing technical difficulties.

The next three lemmas pick out special cases of estimates in the moving cone lemma that will be used in the proof that the graph transform is a contraction mapping (see Proposition 4.11).

Lemma 4.3. *Suppose that $a_0 = 0$. Then there is $\varepsilon_* > 0$ so that for all $\varepsilon \in [0,\varepsilon_*]$,*

$$b(T) \leq b_0/16.$$

Proof. The inequality (4.20) with $t = T$, and $a_0 = 0$, reduces to

$$|a|_{\beta/2,T} \leq C_2\varepsilon T e^{(\beta/2 + C_1\varepsilon)T}|b|_{\beta/2,T},$$

and inserting this in (4.21) we have

$$\begin{aligned}|b|_{\beta/2,T} &\leq b_0 + \varepsilon C_2 C_3 T^2 e^{(\beta/2+C_1\varepsilon)T}|b|_{\beta/2,T}\\ &\leq b_0 + \frac{1}{2}|b|_{\beta/2,T}\end{aligned}$$

by the choice of ε_* in Lemma 4.2 (see (4.23)). Thus $|b|_{\beta/2,T} \leq 2b_0$. But then

$$b(T) \leq e^{-\beta T/2}|b|_{\beta/2,T} \leq \frac{2}{32}b_0 = \frac{1}{16}b_0. \qquad \square$$

LEMMA 4.4. *Suppose $b_0 = 0$. There is $\varepsilon_* > 0$ such that for all $\varepsilon \in [0, \varepsilon_*]$,*

$$b(T) \leq \left(\frac{3L}{4}\right)^2 a_0.$$

Proof. A calculation similar to that above reveals that $b(T) \leq 2C_3 T e^{C_1\varepsilon T} a_0$. When $\varepsilon = 0$ this reduces to

$$b(T) \leq 2C_3 T a_0 \leq \frac{L^2}{4} a_0$$

by the choice of L^2 in Remark 4.6. Thus, for sufficiently small ε_* we obtain the required estimate. $\square$

We conclude with a backward time estimate.

LEMMA 4.5. *Suppose $a(T) = 0$. There is $\varepsilon_* > 0$ such that for all $\varepsilon \in [0, \varepsilon_*]$*

$$a_0 \leq \frac{1}{4L^2} b(T).$$

Proof. We use (4.15) with $t_1 = t$ and $t_2 = T$ to find

$$\begin{aligned}a(t) &\leq e^{C_1\varepsilon(T-t)}a(T) + C_2\varepsilon \int_t^T e^{C_1\varepsilon(s-t)}b(s)ds\\ &= C_2\varepsilon \int_t^T e^{C_1\varepsilon(s-t)}b(s)ds\end{aligned}$$

by our hypothesis. We multiply by $e^{\beta t/2}$ and take the sup over $t \in [0, T]$ to obtain

$$|a|_{\beta/2,T} \leq \frac{C_2\varepsilon}{\beta/2 - C_1\varepsilon}|b|_{\beta/2,T}.$$

Similarly by (4.17), the backward time estimate for $b(t)$ is

$$b(t) \leq e^{3\tilde{\beta}(T-t)}b(T) + C_4 \int_t^T e^{3\tilde{\beta}(s-t)}a(s)ds.$$

We multiply by $e^{\beta t/2}$ and take the sup in t to obtain

$$\begin{aligned}|b|_{\beta/2,T} &\leq e^{3\tilde{\beta}T}\left(b(T) + \frac{C_4 e^{-\beta T/2}}{3\tilde{\beta} - \beta/2}|a|_{\beta/2,T}\right)\\ &\leq e^{3\tilde{\beta}T}\left(b(T) + \frac{C_4 e^{-\beta T/2}}{3\tilde{\beta} - \beta/2}\frac{C_2\varepsilon}{\beta/2 - C_1\varepsilon}|b|_{\beta/2,T}\right).\end{aligned}$$

Let ε_* be so small that for all $\varepsilon \in [0, \varepsilon_*]$,

$$\varepsilon \frac{C_4}{3\tilde{\beta} - \beta/2} \frac{C_2}{2\beta - C_1\varepsilon} e^{(3\tilde{\beta}-\beta/2)T} \leq \frac{1}{2}.$$

Then $|b|_{\beta/2,T} \leq 2e^{3\tilde{\beta}T} b(T)$, and hence

$$a_0 \leq |a|_{\beta/2,T} \leq \varepsilon \frac{2C_2 e^{3\tilde{\beta}T}}{\beta/2 - C_1\varepsilon} b(T).$$

We further reduce ε_* if necessary to obtain $a_0 \leq b(T)/4L^2$ for all $\varepsilon \in [0, \varepsilon_*]$. □

4.5. The graph transform. Define the metric space

$$\mathcal{S}_L = \left\{ h : \Pi_1(\mathcal{D}) \to \mathbb{X}_2 | \ \operatorname{Lip}(h) \leq L, \sup_{u \in \Pi_1(\mathcal{D})} \|h(u)\|_{\mathbb{X}_2} \leq 2R_2 \right\}$$

with the distance function

$$d(h_1, h_2) = \sup_{u \in \Pi_1(\mathcal{D})} \|h_1(u) - h_2(u)\|_{\mathbb{X}_2}.$$

$\mathcal{S}_L$ is complete in this metric. We show below that for any $h \in \mathcal{S}_L$, the image of graph (h) under $\varphi_t, t \in [0, T]$, is the graph of a function in $\mathcal{S}_L$. Taking $t = T$, we define the *graph transform* $\mathcal{G} : \mathcal{S}_L \to \mathcal{S}_L$ by graph $(\mathcal{G}(h)) = \varphi_T$(graph (h)). Most of this subsection is devoted to showing that this definition is unambiguous.

Proposition 4.7 (uniqueness). *Fix $h \in S_L$ and a point $u \in \Pi_1(\mathcal{D})$. There is at most one preimage $u_0 \in \Pi_1(\mathcal{D})$ so that $\Pi_1(\varphi_t(u_0, h(u_0))) = u$.*

Proof. Suppose that $u_1 \neq u_2$ but $\Pi_1(\varphi_t(u_1, h(u_1)) - \varphi_t(u_2, h(u_2))) = 0$. Since $\operatorname{Lip}(h) \leq L$, the point $(u_2, h(u_2))$ lies in the cone $K_L(u_1, h(u_1))$. By the moving cone lemma $\varphi_t(u_2, h(u_2)) \in K_L(\varphi_t(u_1, h(u_1)))$. But then $\Pi_1(\varphi_t(u_1, h(u_1)) - \varphi_t(u_2, h(u_2))) \neq 0$. □

To prove the existence of at least one preimage requires more effort. If $\Pi_1(\mathcal{D})$ were finite-dimensional one could use topological arguments based on degree and the Wazewski principle to prove existence (see, e.g., [3]). This approach would fail here since the manifold to be constructed has both infinite dimension and infinite codimension. Moreover, though we know that there is a solution for $\varepsilon = 0$, we cannot use an implicit function theorem (e.g., as in Fenichel's work [11]) to establish existence for $\varepsilon > 0$ since the perturbation is not Lipschitz in ε. We resort to an explicit solution of the modified equations (4.3)–(4.5) in backward time.

Let $u_T \in \Pi_1(\mathcal{D})$ be fixed. We will show that there exists $(u_0, h(u_0)) \in \mathcal{D}$ such that $\Pi_1(\varphi_T(u_0, h(u_0))) = u_T$. We will rewrite the modified differential equations (4.3)–(4.5) as integral equations in a form different from the mild formulation (4.7)–(4.9). The motivation for this will be clear in the consequent estimates.

Let $S(t, s; u_T)$, $t, s \in \mathbb{R}$, be the two-parameter family in $L(\mathbb{X}_2, \mathbb{X}_2)$ defined as the solution operator to the following linear nonautonomous differential equation:

$$(4.32) \quad \begin{pmatrix} \operatorname{Re}(v)_t \\ \operatorname{Im}(v)_t \\ w_t \end{pmatrix} = \begin{pmatrix} -1 & \delta & \mu \operatorname{Re} f_1(t) \\ -\delta & -1 & \mu \operatorname{Im} f_1(t) \\ -\mu \operatorname{Re} f_1(t) & -\mu \operatorname{Im} f_1(t) & -\gamma_\| \end{pmatrix} \begin{pmatrix} \operatorname{Re}(v) \\ \operatorname{Im}(v) \\ w \end{pmatrix},$$

where

$$f_1(t) = f(e^{\varepsilon(T-t)\partial_x} u_T),$$

and f is defined in (4.2). $S(t,s;u_T)$ is well defined since the right-hand side is a bounded linear operator, and we have the a priori estimate

$$|v(t,x)|^2 + |w(t,x)|^2 \leq e^{-2\beta(t-s)}(|v(s,x)|^2 + |w(s,x)|^2),$$

which ensures the existence of global solutions. In fact, this a priori estimate proves the following.

LEMMA 4.6. $\|S(t_1,t_2;u_T)\| \leq e^{-\beta(t_1-t_2)}$ *for each* $u_T \in \mathbb{X}_1$.

Any mild solution to (4.7)–(4.9) that passes through u_T at time T must satisfy the integral equations

$$u(t) = e^{\varepsilon(T-t)\partial_x} u_T - \varepsilon \int_t^T e^{\varepsilon(s-t)\partial_x} g(u(s),v(s),w(s))ds, \tag{4.33}$$

$$\begin{aligned} \begin{pmatrix} v(t) \\ w(t) \end{pmatrix} = S(t,0;u_T) \begin{pmatrix} v(0) \\ w(0) \end{pmatrix} \\ + \int_0^t S(t,s;u_T) \begin{pmatrix} 0 \\ \gamma_{\|}(\lambda+1) \end{pmatrix} ds \\ +\mu \int_0^t S(t,s;u_T) F(u(s)) \begin{pmatrix} v(s) \\ w(s) \end{pmatrix} ds, \end{aligned} \tag{4.34}$$

where $F(u(s))$ is a skew-symmetric multiplication operator in $L(\mathbb{X}_2, \mathbb{X}_2)$ whose only nonzero terms are

$$\begin{aligned} F_{13} = -F_{31} = \mathrm{Re}(f(u(s))) - f(e^{\varepsilon(T-s)\partial_x} u_T), \\ F_{23} = -F_{32} = \mathrm{Im}(f(u(s))) - f(e^{\varepsilon(T-s)\partial_x} u_T) \end{aligned} \tag{4.35}$$

(we split v into its real and imaginary parts). The virtue of rewriting the equations in this form is that the nonlinear terms are now small. More precisely, by Lemma 4.1 and (4.33)–(4.34), the norm of F is bounded by

$$\begin{aligned} \sup_{s\in[0,T]} \|F(u(s))\| = \sup_{s\in[0,T]} \|f(u(s)) - f(e^{\varepsilon(T-s)\partial_x} u_T)\| \\ \leq 5 \sup_{s\in[0,T]} \|u(s) - e^{\varepsilon(T-s)\partial_x} u_T\| \leq \varepsilon C_5 T \end{aligned} \tag{4.36}$$

for a constant $C_5 = \sup \|g(u,v,w)\| = C_5(\mu,\kappa,R_i)$.

PROPOSITION 4.8 (existence). *There is* $\varepsilon_* > 0$ *such that for each* $\varepsilon \in [0,\varepsilon_*]$ *there exists* $u_0 \in \Pi_1(\mathcal{D})$ *with* $\Pi_1(\varphi_T(u_0,h(u_0))) = u_T$.

Proof. If a preimage exists it must lie in $\Pi_1(\mathcal{D})$ by Remark 4.2. To prove the existence of such a preimage we use iteration on the integral equations (4.33) with the additional condition $(v,w)(0) = h(u(0))$.

Let $u^0(t) = 0$ and $(v,w)^0(t) = 0$ for $0 \leq t \leq T$. For $n \geq 0$ we define the sequence of iterates

$$u^{n+1}(t) = e^{\varepsilon(T-t)\partial_x} u_T - \varepsilon \int_t^T e^{\varepsilon(s-t)\partial_x} g(u^n(s),v^n(s),w^n(s))ds, \tag{4.37}$$

$$\begin{aligned} \begin{pmatrix} v^{n+1}(t) \\ w^{n+1}(t) \end{pmatrix} = S(t,T;u_T) h(u^{n+1}(0)) \\ + \int_0^t S(t,s;u_T) \begin{pmatrix} 0 \\ \gamma_{\|}(\lambda+1) \end{pmatrix} ds \\ + \int_0^t S(t,s;u_T) F(u^{n+1}(s)) \begin{pmatrix} v^n(s) \\ w^n(s) \end{pmatrix} ds. \end{aligned} \tag{4.38}$$

Notice that we solve (4.37) before (4.38).

The sequence defined above satisfies some uniform bounds. First, it is clear that

$$\|u^{n+1}(t) - e^{\varepsilon(T-t)\partial_x} u_T\|_{0,T} \leq \varepsilon T \sup \|g(u,v,w)\| = \varepsilon C_5 T. \tag{4.39}$$

Thus, by Lemmas 4.1, 4.6, and (4.36), we have

$$\|(v,w)^{n+1}(t)\| \leq e^{\beta(T-t)}\|h\| + \frac{(1-e^{-\beta t})}{\beta}(\lambda+1) + 5\varepsilon C_5 T \int_0^t e^{-\beta(t-s)}\|(v,w)^n(s)\| ds.$$

Now $\|h\| \leq 2R_2$ since $h \in \mathcal{S}_L$, so reducing ε_* further if necessary we have

$$\|(v,w)^{n+1}\|_{0,T} \leq \frac{1}{2}\|(v,w)^n\|_{0,T} + \frac{(\lambda+1)}{\beta} + 2R_2 e^{\beta T},$$

where $\|\cdot\|_{0,T} = \sup_{t\in[0,T]}\|\cdot\|$. This implies the uniform bound

$$\sup_{n\geq 0}\|(v,w)^n\|_{0,T} \leq C_6(\beta,\mu,\kappa,\lambda,R_i,T). \tag{4.40}$$

Next, we note that by (4.37) and Lemma 4.1, the difference between consequent iterates of u must satisfy

$$\|u^{n+1} - u^n\|_{0,T} \leq \varepsilon C_7 T(\|u^n - u^{n-1}\|_{0,T} + \|(v,w)^n - (v,w)^{n-1}\|_{0,T}). \tag{4.41}$$

We will estimate each term in the difference between $(v,w)^{n+1}$ and $(v,w)^n$ separately (see (4.38)). The first term is controlled by the uniform Lipschitz constant L.

$$\begin{aligned} &\sup_{t\in[0,T]} \|S(t,T;u_T)(h(u^{n+1}(0)) - h(u^n(0)))\| \\ &\leq \sup_{t\in[0,T]} \|S(t,T;u_T)\|\|h(u^{n+1}(0)) - h(u^n(0))\| \\ &\leq \sup_{t\in[0,T]} e^{-\beta(t-T)} L\|u^{n+1}(0) - u^n(0)\| \leq e^{\beta T} L\|u^{n+1} - u^n\|_{0,T} \\ &\leq \varepsilon C_7 T e^{\beta T} L(\|u^n - u^{n-1}\|_{0,T} + \|(v,w)^n - (v,w)^{n-1}\|_{0,T}) \end{aligned} \tag{4.42}$$

by Lemma 4.6 and inequality (4.41). The differences between the terms on the second line of (4.38) cancel, and the differences between the integrands in the third line are estimated as follows:

$$\begin{aligned} &\left\| F(u^{n+1}(s)) \begin{pmatrix} v^n(s) \\ w^n(s) \end{pmatrix} - F(u^n(s)) \begin{pmatrix} v^{n-1}(s) \\ w^{n-1}(s) \end{pmatrix} \right\| \\ &\leq \|F(u^n)\|\|(v,w)^n - (v,w)^{n-1}\| + \|(v,w)^n\|\|F(u^{n+1}) - F(u^n)\| \\ &\leq \varepsilon C_5 T\|(v,w)^n(s) - (v,w)^{n-1}(s)\| + 5C_6\|u^{n+1}(s) - u^n(s)\|. \end{aligned}$$

In the last step we have used Lemma 4.1 and the uniform estimates (4.36) and (4.40). These terms estimate the integrands. Take the sup over $t \in [0,T]$ and combine the resulting inequality with (4.41) and (4.42) to conclude that the difference between two iterates in (v,w) must satisfy

$$\|(v,w)^{n+1} - (v,w)^n\|_{0,T} \leq \varepsilon C_8 T(\|u^n - u^{n-1}\|_{0,T} + \|(v,w)^n - (v,w)^{n-1}\|_{0,T}). \tag{4.43}$$

We choose ε_* so small that $\max(C_7, C_8)\varepsilon_* T < 1/2$. Then the sequence of iterates is a contraction in the Banach space $C([0,T];\mathbb{X})$. The limit is a trajectory $(u,v,w)(t)$ with $u(T) = u_T$ and $(v,w)(0) = h(u(0))$. □

Remark 4.9. A closer look reveals that we have not used the condition that T is large anywhere in the proof. Thus we have in fact established the stronger statement that for fixed u_T and any $t \in [0,T]$, there is a preimage u_0 so that $\Pi_1(\varphi_t(u_0, h(u_0))) = u_T$. Since u_0 is obtained from a contraction mapping, a slight variant of this argument may be used to prove the existence and uniqueness simultaneously, providing another proof of Proposition 4.7 without invoking Lemma 4.2 (the moving cone lemma). However, the moving cone lemma is of independent interest, and as the proof of Proposition 4.7 shows, it directly implies uniqueness of the preimage.

In essence, the proof reduces to compensating for the unbounded perturbation by viewing the equation in a rotating frame. However, despite the direct proof, the proposition is not trivial. The perturbation is not small but the argument works since the unbounded part of the perturbation generates a unitary group.

We are now in a position to conclude that the graph transform is well defined as a map from $\mathcal{S}_L$ into itself.

Corollary 4.10. $\mathcal{G} : \mathcal{S}_L \to \mathcal{S}_L$.

Proof. Proposition 4.7 and Proposition 4.8 prove that the image of a Lipschitz graph in $\mathcal{S}_L$ is a graph. That the image is also Lipschitz, with Lipschitz constant L, follows from the moving cone lemma. Finally, since $\mathcal{D}$ is positively invariant, the image must satisfy $\|\mathcal{G}(h)\| \le 2R_2$. Thus the graph transform is well defined. □

Now we establish that the graph transform is a contraction mapping on $\mathcal{S}_L$.

Proposition 4.11. *For $\varepsilon \in [0, \varepsilon_*]$ the graph transform $\mathcal{G} : \mathcal{S}_L \to \mathcal{S}_L$ is a contraction.*

Proof. Let $h_i \in \mathcal{S}_L$, $i = 1, 2$. We will show that $d(\mathcal{G}(h_1), \mathcal{G}(h_2)) \le d(h_1, h_2)/2$. Fix u_T in $\Pi_1(\mathcal{D})$. Let $(u_i, h_i(u_i))$ be the unique preimages of $(u_T, \mathcal{G}(h_i)(u_T))$. The distance

$$
\begin{aligned}
& \|\mathcal{G}(h_1)(u_T) - \mathcal{G}(h_2)(u_T)\| \\
& = \|\Pi_2(\varphi_T(u_1, h_1(u_1)) - \varphi_T(u_2, h_2(u_2)))\| \\
& \le \|\Pi_2(\varphi_T(u_1, h_1(u_1)) - \varphi_T(u_1, h_2(u_1)))\| && (4.44)\\
& \qquad + \|\Pi_2(\varphi_T(u_1, h_2(u_1)) - \varphi_T(u_2, h_2(u_1)))\| && (4.45)\\
& \qquad + \|\Pi_2(\varphi_T(u_2, h_2(u_1)) - \varphi_T(u_2, h_2(u_2)))\|. && (4.46)
\end{aligned}
$$

By Lemma 4.3, (4.44) $\le \|h_1(u_1) - h_2(u_1)\|/4$. By Lemma 4.4, (4.45) $\le 3L\|u_1 - u_2\|/4$. And by Lemma 4.3, (4.46) $\le \|h_2(u_1) - h_2(u_2)\|/4 \le L/4\|u_1 - u_2\|$. Thus,

$$
\begin{aligned}
\|\mathcal{G}(h_1)(u_T) - \mathcal{G}(h_2)(u_T)\| &\le \frac{1}{4}\|h_1(u_1) - h_2(u_1)\| + L\|u_1 - u_2\| \\
&\le \frac{1}{4} d(h_1, h_2) + L\|u_1 - u_2\|.
\end{aligned}
$$

Furthermore, by the backward time estimate in Lemma 4.5

$$\|u_1 - u_2\| \le \frac{1}{2L}\|\mathcal{G}(h_1)(u_T) - \mathcal{G}(h_2)(u_T)\|$$

so that

$$\|\mathcal{G}(h_1)(u_T) - \mathcal{G}(h_2)(u_T)\| \le \frac{1}{2} d(h_1, h_2).$$

Since u_T was arbitrary, the lemma is proved. □

COROLLARY 4.12. *There is a unique solution to $\mathcal{G}(h_\varepsilon) = h_\varepsilon$ in $\mathcal{S}_L$. The graph of h_ε, denoted by $\mathcal{M}_\varepsilon$, is invariant under φ_t.*

Proof. We have established that $\varphi_T(\mathcal{M}_\varepsilon) = \mathcal{M}_\varepsilon$. If $0 < t < T$, then $\varphi_t(\mathcal{M}_\varepsilon)$ is the graph of a Lipschitz map in $\mathcal{S}_L$. This follows from Remark 4.9 and the moving cone lemma (notice that Lemma 4.2 is true for all $t \in [0, T]$). But then $\varphi_{T+t}(\mathcal{M}_\varepsilon) = \varphi_t(\mathcal{M}_\varepsilon)$ so that $\varphi_t(\mathcal{M}_\varepsilon)$ is also a fixed point of $\mathcal{G}$. By uniqueness, it follows that $\varphi_t(\mathcal{M}_\varepsilon) = \mathcal{M}_\varepsilon$. Since φ_t is a diffeomorphism, we must have $\varphi_t(\mathcal{M}_\varepsilon) = \mathcal{M}_\varepsilon$ for all $t \in \mathbb{R}$. □

$\mathcal{M}_\varepsilon$ is a *slow manifold*, as it is given as a graph over the slow variable, u. It is clear that all solutions within $\mathcal{D}_0$ are attracted exponentially fast onto the slow manifold. Indeed, given any point in $\mathcal{D}_0$ that does not lie in $\mathcal{M}_\varepsilon$, we may construct a graph in $\mathcal{S}_L$ that passes through this point. Then Proposition 4.11 shows that this graph is attracted exponentially fast onto $\mathcal{M}_\varepsilon$. In particular, this means that the attractor $\mathcal{A}$ is contained within $\mathcal{M}_\varepsilon$. Hence this construction partially answers the open question in [9] on the existence of an inertial manifold in the Maxwell–Bloch equations in the sense that we significantly simplify the geometry of the flow and prove the existence of a smooth, normally hyperbolic invariant manifold attracting all initial conditions. The answer is only partial since this manifold is infinite dimensional.

5. Smoothness of the invariant manifold.

The smoothness of the slow manifold, $\mathcal{M}_\varepsilon$, is established by differentiating the following functional equation that h must satisfy:

$$h(\Pi_1(\varphi_T(u, h(u)))) = \Pi_2(\varphi_T(u, h(u))), \quad u \in \Pi_1(\mathcal{D}). \tag{5.1}$$

For brevity we let $u_T = \Pi_1(\varphi_T(u, h(u)))$ so that (5.1) may be rewritten as

$$h(u_T) = \Pi_2(\varphi_T(u, h(u))), \quad u \in \Pi_1(\mathcal{D}). \tag{5.2}$$

We differentiate (5.2) to obtain a nonlinear functional equation that the derivative of h must satisfy. We prove the existence of a solution to this equation by a contraction mapping argument.

5.1. Notation.

We use the notation in [12] for differentiation. Let $\mathbb{X}_i, \mathbb{Y}$ be Banach spaces. If $F : \mathbb{X} \to \mathbb{Y}$, then $DF : \mathbb{X} \to L(\mathbb{X}, \mathbb{Y})$. If F is a function of several variables, say, $F : \mathbb{X}_1 \times \cdots \times \mathbb{X}_n \to \mathbb{Y}$, then $DF = (D_1F, \ldots, D_nF)$, where $D_iF : \mathbb{X}_1 \times \cdots \times \mathbb{X}_n \to L(\mathbb{X}_i, \mathbb{Y})$. In the interest of brevity we denote

$$P_i = D_i(\Pi_1 \circ \varphi_T), \quad Q_i = D_i(\Pi_2 \circ \varphi_T), \quad i = 1, 2. \tag{5.3}$$

5.2. C^1 smoothness.

Differentiating (5.2) with respect to u we obtain

$$Dh(u_T)Du_T(u) = Q_1(u, h(u)) + Q_2(u, h(u))Dh(u).$$

Since $u_T = \Pi_1 \circ \varphi_T(u, h(u))$, its derivative is

$$Du_T(u) = P_1(u, h(u)) + P_2(u, h(u))Dh(u).$$

Thus, we obtain the formal expression

$$Dh(u_T) = [Q_1 + Q_2Dh(u)]\,[P_1 + P_2Dh(u)]^{-1} \tag{5.4}$$

for the derivative of h. (Here and henceforth we suppress the arguments of P_i, Q_i to simplify notation.) When $\varepsilon = 0$, the derivatives satisfy $P_1 = \mathrm{Id}$, $P_2 = 0$, and

$\|Q_2\| \le e^{-\beta T}$. Furthermore, $u_T = u$. Thus, in this limit, (5.4) reduces to $Dh(u) = Q_1 + Q_2 Dh(u)$, which has a unique solution since $\|Q_2\|$ is small. This suggests that we use iteration to solve (5.4) for $\varepsilon > 0$.

We now define the function space in which we wish to construct the derivative Dh. Let

$$\mathcal{T}_L = \left\{ A : \Pi_1(\mathcal{D}) \to L(\mathbb{X}_1, \mathbb{X}_2) | \sup_{u \in \Pi_1(\mathcal{D})} \|A(u)\|_{L(\mathbb{X}_1,\mathbb{X}_2)} \le L \right\}$$

be the metric space of continuous maps with the distance function

$$d(A_1, A_2) = \sup_{u \in \Pi_1(\mathcal{D})} \|A_1(u) - A_2(u)\|_{L(\mathbb{X}_1,\mathbb{X}_2)}.$$

$\mathcal{T}_L$ is complete in this metric.

We also define a map $\mathcal{F} : \mathcal{T}_L \to \mathcal{T}_L$ as

$$\mathcal{F}(A)(u_T) = [Q_1 + Q_2 A(u)] [P_1 + P_2 A(u)]^{-1}. \tag{5.5}$$

We shall prove that $\mathcal{F}$ is a contraction and the unique fixed point $\mathcal{F}(A) = A$ is the derivative of h. This will imply that $\mathcal{M}_\varepsilon$ is at least of class C^1.

We will use the following lemmas to estimate the terms in (5.5).

Lemma 5.1. *There is $\varepsilon_* > 0$ so that for $\varepsilon \in [0, \varepsilon_*]$*

(a) $\sup_{u\in\Pi_1(\mathcal{D})} \|P_1 - e^{-\varepsilon T \partial_x}\|_{L(\mathbb{X}_1,\mathbb{X}_1)} = O(\varepsilon)$,

(b) $\sup_{u\in\Pi_1(\mathcal{D})} \|P_2\|_{L(\mathbb{X}_1,\mathbb{X}_2)} = O(\varepsilon)$,

(c) $\sup_{u\in\Pi_1(\mathcal{D})} \|Q_1\|_{L(\mathbb{X}_2,\mathbb{X}_1)} \le L/4$,

(d) $\sup_{u\in\Pi_1(\mathcal{D})} \|Q_2\|_{L(\mathbb{X}_2,\mathbb{X}_2)} \le 1/8$.

Proof. The proof entails estimating the growth of derivatives using the equation of variations. The arguments are direct but tedious so we will omit a few details. The main point is that despite the singular perturbation, we can control the derivatives with knowledge of the limit $\varepsilon = 0$ provided we account for the unbounded terms properly (e.g., as in statement (a) of the lemma).

We start by redefining $S(t, s; u_0)$, $t, s \in \mathbf{R}$, as the solution operator to the linear nonautonomous differential equation (4.32) with

$$f_1(t) = f(e^{-\varepsilon t \partial_x} u_0). \tag{5.6}$$

Notice that Lemma 4.6 remains valid with this definition of f_1. The mild formulation is the obvious analogue of (4.33)–(4.34) provided we redefine $F(u)$ as the skew-symmetric multiplication operator whose only nonzero terms are

$$\begin{aligned} F_{13} &= -F_{31} = \mathrm{Re}(f(u(s)) - f(e^{-\varepsilon s \partial_x} u_0)), \\ F_{23} &= -F_{32} = \mathrm{Im}(f(u(s)) - f(e^{-\varepsilon s \partial_x} u_0)). \end{aligned} \tag{5.7}$$

The estimate (4.36) shows that $F(u(s))$ is uniformly small on $[0, T]$. Differentiating (4.33) with respect to the initial point u_0, we obtain linear integral equations that the derivatives must satisfy.

$$\begin{aligned} D_{u_0} u(t) = e^{-\varepsilon t \partial_x} &+ \varepsilon \int_0^t e^{-\varepsilon(t-s)\partial_x} \left(D_1 g D_{u_0} u(s) + D_2 g D_{u_0} v(s) \right) ds \\ &+ \varepsilon \int_0^t e^{-\varepsilon(t-s)\partial_x} D_3 g D_{u_0} w(s) ds \end{aligned} \tag{5.8}$$

and

$$\begin{aligned}\begin{pmatrix} D_{u_0}v(t) \\ D_{u_0}w(t) \end{pmatrix} &= D_{u_0}S(t,0;u_0)\begin{pmatrix} v_0 \\ w_0 \end{pmatrix} \\ &\quad + \int_0^t D_{u_0}(S(t,s;u_0))\left(\begin{pmatrix} 0 \\ \lambda+1 \end{pmatrix} + F(u(s))\begin{pmatrix} v(s) \\ w(s) \end{pmatrix}\right) ds \\ &\quad + \int_0^t S(t,s;u_0)DF(u(s))D_{u_0}u(s)\begin{pmatrix} v(s) \\ w(s) \end{pmatrix} ds \\ &\quad + \int_0^t S(t,s;u_0)F(u(s))\begin{pmatrix} D_{u_0}v(s) \\ D_{u_0}w(s) \end{pmatrix} ds. \end{aligned} \tag{5.9}$$

The derivative of $S(t,0;u_0)$ is computed from its definition in (4.32). Since $S(t,0;u_0)$ is defined by the solution to a system of linear nonautonomous equations, its derivative $D_{u_0}S(t,0;u_0)$ is a linear map from $\mathbb{X}_1 \to L(\mathbb{X}_2,\mathbb{X}_2)$ defined for any $u_1 \in \mathbb{X}_1$ by

$$D_{u_0}S(t,0;u_0)u_1 = \int_0^t S(t,s;u_0)(D_{u_0}G(s,0;u_0)u_1)S(s,0;u_0)ds,$$

where $G(t,0;u_0)$ is the matrix defined on the right-hand side of (4.32) with $f_1(t)$ redefined as in (5.6). It follows from Lemma 4.1 and Lemma 4.6 that $\|D_{u_0}S(t,0;u_0)\| \leq 5te^{-\beta t}$. Thus the "linear" part of $D_{u_0}(v(t),w(t))$ is bounded for all u_0 and for all $t \in [0,T]$ by some constant C_9. By the choice of L (see Remark 4.6) $C_9 \leq L/8$. The nonlinear part in (5.9) (i.e., the terms with F and DF) are $O(\varepsilon)$ by Lemma 4.6 and the estimate (4.36). The nonlinear terms in (5.8) are also $O(\varepsilon)$ since $D_i g$ is uniformly bounded (see Lemma 4.1). Thus, one may prove that a solution to the equation of variations (5.8) and (5.9) exists for sufficiently small ε_* by a contraction mapping argument as in the proof of Proposition 4.8. Then Gronwall estimates show that $\sup_{t\in[0,T]} \max(\|D_{u_0}u(t)\|, \|D_{u_0}(v(t),w(t))\|) \leq C(T,Q_i)$ for all $(u_0,v_0,w_0) \in \mathcal{D}$ so that for all $\varepsilon \in [0,\varepsilon_*]$,

$$\|D_{u_0}u(T) - e^{-\varepsilon T\partial_x}\|_{L(\mathbb{X}_1,\mathbb{X}_1)} \leq \varepsilon C(T,R_i).$$

This proves (a). Similarly, the deviation of Q_1 from its linear part is $O(\varepsilon)$ and for small ε we have (c). Estimates (b) and (d) are obtained from the equation of variations for the derivative in (v_0,w_0). These are

$$\begin{aligned} D_{(v_0,w_0)}u(t) &= \varepsilon\int_0^t e^{-\varepsilon(t-s)\partial_x}\left[D_1gD_{(v_0,w_0)}u(s) + D_2gD_{(v_0,w_0)}v(s)\right] ds \\ &\quad + \varepsilon\int_0^t e^{-\varepsilon(t-s)\partial_x}D_3gD_{(v_0,w_0)}w(s)ds, \\ \begin{pmatrix} D_{(v_0,w_0)}v(t) \\ D_{(v_0,w_0)}w(t) \end{pmatrix} &= S(t,0;u_0) + \int_0^t S(t,s;u_0)DF(u(s))D_{(v_0,w_0)}u(s)ds \\ &\quad + \int_0^t S(t,s;u_0)F(u(s))\begin{pmatrix} D_{(v_0,w_0)}v(s) \\ D_{(v_0,w_0)}w(s) \end{pmatrix} ds. \end{aligned}$$

Again, the nonlinear terms are $O(\varepsilon)$ so these equations can be solved by a contraction mapping argument. Gronwall estimates show that $\|D_{(v_0,w_0)}u(t)\|$ and $\|D_{(v_0,w_0)}(v(t), w(t)) - S(t,0;u_0)\|$ are $O(\varepsilon)$ with constants that depend only on R_i and T. But $\|S(t,0;u_0)\|$ is exponentially decaying by Lemma 4.6, and since T has been chosen so large that $e^{-\beta T/2} = 1/32$, we may further reduce ε_* to obtain (b) and (d). □

LEMMA 5.2. *There is $\varepsilon_* > 0$ so that for each $\varepsilon \in [0, \varepsilon_*]$ and $A \in \mathcal{T}_L$ we have*

$$\sup_{u \in \Pi_1(\mathcal{D})} \|[P_1 + P_2 A(u)]^{-1}\|_{L(\mathbb{X}_1, \mathbb{X}_1)} \leq 1 + O(\varepsilon) \leq 2.$$

Proof. We fix $u \in \Pi_1(\mathcal{D})$ and write

$$\begin{aligned}[P_1 + P_2 A]^{-1} &= [e^{-\varepsilon T \partial_x} - (e^{-\varepsilon T \partial_x} - P_1 - P_2 A)]^{-1} \\ &= e^{\varepsilon T \partial_x} [\mathrm{Id} - e^{\varepsilon T \partial_x} (e^{-\varepsilon T \partial_x} - P_1 - P_2 A)]^{-1}.\end{aligned}$$

This suggests that we write the inverse as a Neumann series. If $A \in \mathcal{T}_L$, then its norm is bounded by L. Thus by Lemma 5.1,

$$\|e^{-\varepsilon T \partial_x} - P_1 + P_2 A(u)\| \leq \|e^{-\varepsilon T \partial_x} - P_1\| + L\|P_2\| \leq C\varepsilon. \tag{5.10}$$

Also note that $e^{\varepsilon T \partial_x}$ is an isometry on $\mathbb{X}_1$. Thus, the Neumann series converges for $[P_1 + P_2 A]^{-1}$ for ε_* sufficiently small, and the norm of the sum does not exceed $1 + O(\varepsilon)$, which for sufficiently small ε is less than 2. □

Let $A \in \mathcal{T}_L$. Then we obtain

$$\|\mathcal{F}(A)\| \leq (\|Q_1\| + L\|Q_2\|)(1 + O(\varepsilon))$$

by the previous lemma. Furthermore, by Lemma 5.1, $\|Q_1\| + \|Q_2\| L \leq L/2$. Thus, for ε_* sufficiently small $\|\mathcal{F}(A)\| \leq L$, and hence $\mathcal{F}$ is well defined. The next proposition shows that for sufficiently small ε_*, it is in fact a contraction.

PROPOSITION 5.1. *There is $\varepsilon_* > 0$ such that for $\varepsilon \in [0, \varepsilon_*]$ the mapping $\mathcal{F} : \mathcal{T}_L \to \mathcal{T}_L$ is a contraction.*

Proof. We let $A, B \in \mathcal{T}_L$, fix $u \in \Pi_1(\mathcal{D})$, and let $u_T = \Pi_1 \circ \varphi_T(u, h(u))$. Then

$$\begin{aligned}&\mathcal{F}(A)(u_T) - \mathcal{F}(B)(u_T) = Q_2(A(u) - B(u))[P_1 + P_2 A(u)]^{-1} \\ &+(Q_1 + Q_2 B(u)) \left([P_1 + P_2 A(u)]^{-1} - [P_1 + P_2 B(u)]^{-1}\right).\end{aligned} \tag{5.11}$$

Applying Lemmas 5.1 and 5.2 to the first term we have

$$\|Q_2(A(u) - B(u))[P_1 + P_2 A(u)]^{-1}\| \leq \frac{2}{8}\|A(u) - B(u)\|. \tag{5.12}$$

We use the identity $(M - A)^{-1} - (M - B)^{-1} = (M - A)^{-1}(A - B)(M - B)^{-1}$ and Lemma 5.2 to estimate the second term in (5.11)

$$\begin{aligned}&(Q_1 + Q_2 B(u)) \left([P_1 + P_2 A(u)]^{-1} - [P_1 + P_2 B(u)]^{-1}\right) \\ &\leq 4(\|Q_1\| + L\|Q_2\|)\|P_2\|\|A(u) - B(u)\| \leq 2L\|P_2\|\|A(u) - B(u)\| \\ &\leq C\varepsilon\|A(u) - B(u)\| \leq \frac{1}{2}\|A(u) - B(u)\|\end{aligned} \tag{5.13}$$

for sufficiently small ε_*. Thus, $\|\mathcal{F}(A)(u_T) - \mathcal{F}(B)(u_T)\| \leq 3/4\|A(u) - B(u)\|$. Since u was arbitrary, this proves the lemma. □

To complete the proof that $\mathcal{M}_\varepsilon$ is C^1, we must show that the unique fixed point of $\mathcal{F}$ is indeed the derivative Dh. This step is essentially the same as Proposition 7 in Fenichel's paper [11], so the proof is omitted.

5.3. C^k smoothness. Higher order smoothness will be proven using the following bootstrapping argument of Fenichel [11]. The unique fixed point $A = \mathcal{F}(A)$ can be realized as the limit of a sequence of iterates $A^n = \mathcal{F}^n A^0$ with $A^0 = 0$. For any $u \in \Pi_1(\mathcal{D})$,

$$A^{n+1}(u_T) = [Q_1 + Q_2 A^n(u)][P_1 + P_2 A^n(u)]^{-1}. \tag{5.14}$$

Since h is C^1, the maps $P_i = P_1(u, h(u))$ and Q_i are differentiable, so A^{n+1} is differentiable if A^n is. Thus, to show that the limit A is differentiable it suffices to show that the sequence $\{DA^n\}$ converges in the space $C(\Pi_1(\mathcal{D}), L^2(\mathbb{X}_1, \mathbb{X}_2))$.

We will show this with estimates similar to those of Proposition 5.1. From the proof of Proposition 5.1 (in particular, (5.12) and (5.13) with $A^{n-1} = A$ and $A^n = B$) we see that the principal term in the contraction estimate at the nth step is

$$\begin{aligned} &\|(P_1 + P_2 A^{n-1})^{-1}\| \left(\|Q_2\| + \|P_2\| \|Q_1 + Q_2 A^n\| \|(P_1 + P_2 A^n)^{-1}\|\right) \\ &\quad \le (1 + C\varepsilon)\left(\frac{1}{8} + C\varepsilon\right) := \alpha_1 \end{aligned} \tag{5.15}$$

by Lemma 5.1 and Lemma 5.2. Higher order derivatives can be obtained in the same way. Differentiating (5.14) we obtain

$$\begin{aligned} DA^{n+1}(u_T) Du_T(u) &= Q_2 DA^n(u)[P_1 + P_2 A^n(u)]^{-1} \\ &-[Q_1 + Q_2 A^n][P_1 + P_2 A^n]^{-1} P_2 DA^n [P_1 + P_2 A^n]^{-1} + \text{lower order terms}, \end{aligned} \tag{5.16}$$

where the lower order terms do not involve derivatives in A. Thus, the principal term in the contraction estimate is now

$$\begin{aligned} &\left(\|Q_2\| + \|P_2\| \|Q_1 + Q_2 A^n\| \|(P_1 + P_2 A^n)^{-1}\|\right) \|(P_1 + P_2 A^{n-1})^{-1}\|^2 \\ &\quad \le (1 + C\varepsilon)^2 \left(\frac{1}{8} + C\varepsilon\right) := \alpha_2. \end{aligned} \tag{5.17}$$

For ε_* sufficiently small, $\alpha_2 < 1$ holds for $0 \le \varepsilon \le \varepsilon_*$. Let $a_n = \sup_u \|DA^{n+1}(u_T) - DA^n(u_T)\|$. It follows from (5.16) and (5.17) that

$$a_{n+1} \le \alpha_2 a_n + r_n,$$

where r_n is a remainder term obtained from the differences in lower order terms. r_n diminishes to zero as n increases since A^n converges. Thus, for any $\eta > 0$ there exists an N such that $r_n \le \eta$ for all $n \ge N$. Hence,

$$a_{N+m} \le \alpha_2^m a_N + \frac{\eta}{1 - \alpha_2},$$

and thus $\limsup_{n\to\infty} a_n \le \eta/(1 - \alpha_2)$. Since η was arbitrary, $a_n \to 0$, and it follows that the sequence $\{DA^n\}$ converges. Thus, $\mathcal{M}_\varepsilon$ is of class C^2.

We now proceed inductively. To show that $\mathcal{M}_\varepsilon$ is C^k assuming that it is C^{k-1}, it is sufficient to show that the sequence $\{D^k A^n\}$ converges. Each term in the sequence is of the form

$$\begin{aligned} D^k A^{n+1}(u_T) Du_T(u) &= Q_2 D^k A^n(u)[P_1 + P_2 A^n(u)]^{-k} \\ &-[Q_1 + Q_2 A^n][P_1 + P_2 A^n]^{-1} P_2 D^k A^n [P_1 + P_2 A^n]^{-k} + \text{terms of order } k-1, \end{aligned}$$

and the principal term in the contraction estimate is bounded by

$$(1+C\varepsilon)^k\left(\frac{1}{8}+C\varepsilon\right):=\alpha_k.$$

For $\varepsilon_*(k)$ sufficiently small, $\alpha_k<1$ for all $\varepsilon\in[0,\varepsilon_*]$ and the sequence $\{D^kA^n\}$ is convergent. Thus, $\mathcal{M}_\varepsilon$ is of class C^k. The manifold is not C^∞ since it is clear that $\varepsilon_*(k)$ must decrease to zero as k increases arbitrarily. This completes the proof of the C^k smoothness and the proof of Theorem 4.4.

6. Geometric singular perturbation theory.

6.1. Notation. In this section we need to distinguish carefully between the flow for different values of ε. To emphasize this, we will use the superscript ε. For example, φ_t^ε denotes the flow with a particular choice of ε, $(u^\varepsilon(t),v^\varepsilon(t),w^\varepsilon(t))$ denotes a trajectory, and $\mathcal{A}_\varepsilon$ denotes the attractor for φ_t^ε. We use the same notation for the modified and unmodified flow, but the flow under consideration will be clear from the context.

6.2. Reduced dynamics and the slaving principle. Theorem 4.4 provides a rigorous decomposition of the flow and a justification of the "slaving principle." First consider the modified flow. Since $\mathcal{M}_\varepsilon$ is invariant, any trajectory on it must satisfy

$$u(t)=e^{-\varepsilon t\partial_x}u(0)+\varepsilon\int_0^t e^{-\varepsilon(t-s)\partial_x}g(u(s),h^\varepsilon(u(s)))ds, \tag{6.1}$$

$$h_v^\varepsilon(u(t))=e^{-(1+i\delta)t}h_v^\varepsilon(u(0))+\mu\int_0^t e^{-(1+i\delta)(t-s)}f(u(s))h_w^\varepsilon(u(s))ds, \tag{6.2}$$

$$\begin{aligned} h_w^\varepsilon(u(t)) &= e^{-\gamma_\| t}h_w^\varepsilon(u(0))+(\lambda+1)(1-e^{-\gamma_\| t})\\ &\quad -\mu\int_0^t e^{-(t-s)\gamma_\|}\,\mathrm{Re}\,(f(u(s))^*h_v^\varepsilon(u(s)))ds. \end{aligned} \tag{6.3}$$

Thus, the slow dynamics decouples from the fast dynamics. This is only half the story: we have established the existence of a reduced equation that is a *functional* differential equation, but we have not prescribed a formula to compute the reduced equation. Theorem 4.4 proves the existence of a family of invariant manifolds $\{\mathcal{M}_\varepsilon\}_{\varepsilon\in[0,\varepsilon_*]}$. In Fenichel's theory [12] these manifolds $\mathcal{M}_\varepsilon$ fit together smoothly in ε and there is a global center manifold given as a function $h(\varepsilon,u)=h^\varepsilon(u)$. Thus we may expand $h^\varepsilon(u)=h(0,u)+D_1h(0,u)\varepsilon+R(u,\varepsilon)$, where $R=o(\varepsilon)$. In infinite dimensions the situation is considerably more delicate. The issue is, of course, the unbounded term $\varepsilon\partial_x$. For flows that are close in the C^1 topology, it usually follows from an implicit function theorem, or the proof of the existence of the invariant manifold, that the unperturbed and perturbed manifolds are close. In the presence of unbounded perturbations the convergence of h^ε to h^0 is expressed in the following theorem.

THEOREM 6.1. $\|h^\varepsilon(u)-h^0(u)\|\to 0$ *uniformly on compact sets.*

Proof. The only information we have is that $\mathcal{M}_\varepsilon$ is invariant under the flow φ_t^ε. Thus the proof will rely on the closeness of φ_t^ε to φ_t^0. We fix $u\in\Pi_1(\mathcal{D}_0)$ and let $u(0)=u$. We will estimate the difference $h^\varepsilon(u)-h^0(u)$ by using the integral equations (6.1)–(6.3). When $\varepsilon=0$, (6.1)–(6.3) reduce to the algebraic equations $u(t)\equiv u$ and

$$\begin{pmatrix} 1 & \delta & -\mu\mathrm{Re}f(u)\\ -\delta & 1 & -\mu\mathrm{Im}f(u)\\ \mu\mathrm{Re}f(u) & \mu\mathrm{Im}f(u) & \gamma_\| \end{pmatrix}\begin{pmatrix}\mathrm{Re}\,h_v^0(u)\\ \mathrm{Im}\,h_v^0(u)\\ h_w^0(u)\end{pmatrix}=\begin{pmatrix}0\\0\\ \gamma_\|(\lambda+1)\end{pmatrix}, \tag{6.4}$$

or, more briefly,

$$B(u)h^0(u) = (0, 0, \gamma_{\|}(\lambda + 1))^T, \tag{6.5}$$

where $B(u)$ is the multiplication operator in (6.4). This is the expression that was computed in (3.1). For every $x \in S^1$, the matrix $B(u(x))$ is invertible since

$$\det(B(u(x))) = \gamma_{\|}(1 + \delta^2) + \mu^2 |f(u)(x)|^2 \geq \gamma_{\|}(1 + \delta^2) \tag{6.6}$$

is uniformly bounded away from zero. Thus, $B(u)$ is invertible and its inverse, $B(u)^{-1}$, is the multiplication operator in $L(\mathbb{X}_2, \mathbb{X}_2)$ defined for every $x \in S^1$ by the matrix $B(u(x))^{-1}$. From the algebraic formula for the inverse of a matrix, and the estimate (6.6), we see that

$$\sup_{u \in \Pi_1(\mathcal{D})} \|B(u)^{-1}\| \leq C_B(R_1). \tag{6.7}$$

Define the function $\Psi : \mathbb{X} \to \mathbb{X}_2$ by

$$\Psi(u, v, w) = B(u) \begin{pmatrix} v \\ w \end{pmatrix} - \begin{pmatrix} 0 \\ \gamma_{\|}(\lambda + 1) \end{pmatrix}. \tag{6.8}$$

Notice that $\Psi(u, v, w) = 0$ if and only if $(v, w) = h^0(u)$. Let K be a compact subset of $\Pi_1(\mathcal{D})$. We will show that $\lim_{\varepsilon \downarrow 0} \sup_{u \in K} \|\Psi(u, h^\varepsilon(u))\| = 0$. This implies the theorem since

$$h^\varepsilon(u) - h^0(u) = B(u)^{-1}B(u)h^\varepsilon(u) - B(u)^{-1}(0, \gamma_{\|}(\lambda+1))^T = B(u)^{-1}\Psi(u, h^\varepsilon(u)), \tag{6.9}$$

and hence from (6.7) we obtain

$$\sup_{u \in K} \|h^\varepsilon(u) - h^0(u)\| \leq C_B \sup_{u \in K} \|\Psi(u, h^\varepsilon(u))\|.$$

We will consider the components of Ψ separately. The first two components of Ψ are $(1 + i\delta)h_v^\varepsilon(u) - \mu f(u) h_w^\varepsilon(u)$ (taking the real and imaginary components together). We estimate this term using (6.2). We start with the initial condition $(u, v, w)(0) = (u, h^\varepsilon(u))$ and then calculate that for any $t > 0$,

$$\begin{aligned}(1 - e^{-(1+i\delta)t}) \left[-(1 + i\delta)h_v^\varepsilon(u) + \mu f(u) h_w^\varepsilon(u)\right] &= (1 + i\delta)\Bigg[h_v^\varepsilon(u^\varepsilon(t)) - h_v^\varepsilon(u) \\ -\mu \int_0^t e^{-(1+i\delta)(t-s)} \left(f(u^\varepsilon(s))h_w^\varepsilon(u^\varepsilon(s)) - f(u)h_w^\varepsilon(u)\right) ds\Bigg]\end{aligned} \tag{6.10}$$

(since u is a constant, we can take $\mu f(u) h_w^\varepsilon(u)$ under the integral sign). Notice that

$$\begin{aligned}\mathrm{Lip}(f h_w^\varepsilon) &\leq \left(\sup_u \|f(u)\|\right) \mathrm{Lip}(h_w^\varepsilon) + \left(\sup_u \|h_w^\varepsilon(u)\|\right) \mathrm{Lip}(f) \\ &\leq 2R_1 L + 2R_2 \cdot 5 := C_{10}\end{aligned}$$

by Lemma 4.1 and the definition of $\mathcal{S}_L$. Thus, we obtain from (6.10) that

$$\begin{aligned}&\|(1 + i\delta)h_v^\varepsilon(u) - \mu f(u)h_w^\varepsilon(u)\| \\ &\leq \frac{|1 + i\delta|}{|1 - e^{-(1+i\delta)t}|}\left[L\|u^\varepsilon(t) - u\| + \mu \int_0^t C_{10} e^{-(t-s)} \|u^\varepsilon(s) - u\| ds\right] \\ &\leq C_{11}\left[\frac{\|u^\varepsilon(t) - u\|}{|1 - e^{-(1+i\delta)t}|} + \frac{1 - e^{-t}}{|1 - e^{-(1+i\delta)t}|} \sup_{s \in [0,t]} \|u^\varepsilon(s) - u\|\right].\end{aligned}$$

Notice that $(1-e^{-t})/|1-e^{-(1+i\delta)t}|$ is uniformly bounded for all $t>0$. And for t in any fixed domain $(0,T]$ we have $1/|1-e^{-(1+i\delta)t}| \leq C(T)/t$. Thus we find that

$$(6.11)\quad \|(1+i\delta)h_v^\varepsilon(u)-\mu f(u)h_w^\varepsilon(u)\| \leq C\left[\frac{\|u^\varepsilon(t)-u\|}{t}+\sup_{s\in[0,t]}\|u^\varepsilon(s)-u\|\right].$$

A similar calculation shows that we obtain the same result for the second component of Ψ. Thus, we find

$$(6.12)\qquad \|\Psi(u,h^\varepsilon(u))\| \leq C\left[\frac{\|u^\varepsilon(t)-u\|}{t}+\sup_{s\in[0,t]}\|u^\varepsilon(s)-u\|\right].$$

Finally, we use (6.1) to estimate the difference $u^\varepsilon(t)-u$. The difference consists of two parts, the deviation from the linear part of the flow $e^{-\varepsilon t\partial_x}u$ and the deviation of the linear flow from the nonlinear flow. Precisely,

$$\|u^\varepsilon(t)-u\| \leq \|u^\varepsilon(t)-e^{-\varepsilon t\partial_x}u\|+\|e^{-\varepsilon t\partial_x}u-u\| \leq C\varepsilon t+\|e^{-\varepsilon t\partial_x}u-u\|.$$

Inserting this estimate in (6.12) we have

$$(6.13)\quad \|\Psi(u,h^\varepsilon(u))\| \leq C\left[\frac{\|e^{-\varepsilon t\partial_x}u-u\|}{t}+\sup_{s\in[0,t]}\|e^{-\varepsilon t\partial_x}u-u\|+\varepsilon(1+t)\right].$$

For fixed t and u, the right-hand side of (6.13) goes to zero as $\varepsilon\downarrow 0$. Next suppose that we fix t but consider u ranging over a compact subset K. Since functions in K are equicontinuous, $\sup_{u\in K}\|e^{-\varepsilon t\partial_x}u-u\| \to 0$ as $\varepsilon\downarrow 0$. □

The estimate (6.13) highlights why the convergence of h^ε to h^0 is not any better than uniform convergence on compact subsets. Since t is a free parameter, the estimate is best when we take the infimum with respect to t. Since the flow is continuous in t we must have $\sup_{s\in[0,t]}\|u^\varepsilon(s)-u\| \to 0$ as $t\to 0$. But the first term in (6.12) may not have a limit. The reason is that $\lim_{t\downarrow 0}\|e^{-\varepsilon t\partial_x}u-u\|/t$ does not exist for most functions (in the sense of category). If u is C^1, then we find that

$$\|\Psi(u,h^\varepsilon(u))\| \leq C\left(\|Du\|_\infty\varepsilon+\sup_{s\in[0,t]}\|Du\|_\infty\varepsilon t+\varepsilon(1+t)\right),$$

and since t is a free parameter, we take the infimum over t to find

$$(6.14)\qquad \|\Psi(u,h^\varepsilon(u)) \leq C(\|Du\|_\infty+1)\varepsilon.$$

Another example of more rapid convergence is provided by taking K to be a bounded subset of $C^{0,\alpha}(S^1;\mathbf{C})$, the space of Hölder continuous functions with modulus $\alpha\in(0,1]$. In this case we find that

$$\sup_{u\in K}\|e^{-\varepsilon t\partial_x}u-u\| \leq H\varepsilon^\alpha t^{\alpha-1},$$

where H is the maximum Hölder seminorm of the functions in K. Then we take the infimum in t on both sides of (6.13) to find that

$$\sup_{u\in K}\|h^\varepsilon(u)-h^0(u)\| \leq C\varepsilon^\alpha.$$

6.3. Formal asymptotic expansions. Equations (6.1)–(6.3) are also the starting point for a formal asymptotic expansion. Theorem 6.1 shows that we can control the remainder only if $u(s)$ has some smoothness in x. However, a *formal* asymptotic expansion may be obtained by using the invariance of $\mathcal{M}_\varepsilon$. Make the ansatz

$$h^\varepsilon(u) = h^0(u) + \varepsilon h_1(u) + \varepsilon^2 h_2(u) + \cdots. \tag{6.15}$$

Substituting this ansatz in (6.1)–(6.3) and matching the powers of ε we obtain after some calculations that

$$h_n(u) = c_n(u)\partial_x^n u + d_n(u, \partial_x u, \ldots, \partial_x^{n-1}u), \quad n \geq 1, \tag{6.16}$$

where $c_n(u)(x)$ depends only on $u(x)$. This expansion suggests that $h^\varepsilon(u)(x)$ actually depends on the germ of u at x. Thus, we expect $h^\varepsilon(u)$ to have a nonlocal dependence on u. The expansion also suggests that the reduced equation (6.17) is not hyperbolic because $h_n(u)$ includes higher order diffusive and dispersive terms. In fact, the reduced equation cannot be hyperbolic for if it were, there would be no asymptotic smoothing on the attractor. Similar questions arise in hyperbolic conservation laws with relaxation. We refer especially to the article by Chen, Levermore, and Liu, section 2 of which contains the same geometric description of formal reductions in the context of conservation laws [5].

6.4. Regular dynamics. We can now revert to a description of the unmodified Maxwell–Bloch equations in the slow (and natural) time scale. Changing the time scale to $\tau = \varepsilon t$ we have for all $u(0) \in \Pi_1(\mathcal{D}_0)$ and $\tau \geq 0$,

$$u(\tau) = e^{-\kappa\tau}e^{-\tau\partial_x}u(0) + \frac{\kappa}{\mu}\int_0^\tau e^{-\kappa(\tau-s)}e^{-(\tau-s)\partial_x}h_v^\varepsilon(u(s))ds. \tag{6.17}$$

We have used the positive invariance of $\Pi_1(\mathcal{D}_0)$ and the fact that $g(u,v,w)$ reduces to v within the domain $\mathcal{D}_0$. To make the comparison with the formal reduction precise, we shall write (6.17) as

$$\begin{aligned} u(\tau) = e^{-\kappa\tau}e^{-\tau\partial_x}u(0) &+ \frac{\kappa}{\mu}\int_0^\tau e^{-\kappa(\tau-s)}e^{-(\tau-s)\partial_x}h_v^0(u(s))ds \\ &+ \frac{\kappa}{\mu}\int_0^\tau e^{-\kappa(\tau-s)}e^{-(\tau-s)\partial_x}(h_v^\varepsilon(u(s)) - h_v^0(u(s)))ds. \end{aligned} \tag{6.18}$$

The attractor $\mathcal{A}_\varepsilon$ is an invariant set contained in $\mathcal{M}_\varepsilon$. On the attractor, the reduction is valid uniformly in time. Applying (6.14) to $u \in \Pi_1(\mathcal{A}_\varepsilon)$ we have

$$\begin{aligned} &\left\| u(\tau) - e^{-\kappa\tau}e^{-\tau\partial_x}u(0) - \frac{\kappa}{\mu}\int_0^\tau e^{-\kappa(\tau-s)}e^{-(\tau-s)\partial_x}h_v^0(u(s))\right\| \\ &\leq \frac{\kappa}{\mu}\int_0^\tau e^{-\kappa(\tau-s)}\|h^\varepsilon(u(s)) - h^0(u(s))\|ds \leq C\varepsilon\left(\sup_{u\in\mathcal{A}_\varepsilon}\|Du\|_\infty + 1\right) \end{aligned}$$

for all τ. Unfortunately, this isn't enough as the estimates of section 4 in [9] show that $\sup_{u\in\mathcal{A}_\varepsilon}\|Du\|_\infty = O(1/\varepsilon)$. Furthermore, based on numerical evidence we expect that this estimate is sharp. In several parameter regimes the Lorenz ODEs have periodic solutions with arbitrarily large period. These solutions in turn imply the existence of traveling wave solutions to the Maxwell–Bloch equations with gradients of $O(1/\varepsilon)$. In fact, estimating $\sup_{u\in\mathcal{A}_\varepsilon}\|\partial_x^n u\|_\infty$ we find that the series (6.16) diverges even on the attractor.

6.5. Change of stability under perturbation. The guiding philosophy of geometric singular perturbation for ODEs is that normally hyperbolic manifolds within the formally reduced flow persist for the perturbed flow, provided the critical manifold is normally hyperbolic. As an example of this, Fenichel proved a theorem of Anosov on the persistence of periodic orbits for a singularly perturbed ODE [12]. A simpler example is to consider hyperbolic fixed points. Let the reduced flow have an exponentially attracting fixed point, and let the critical manifold be exponentially attracting. Then Theorem 12.1 in [12] shows that the fixed point persists for $\varepsilon > 0$ and remains attracting.

Even this simple assertion is false for PDE; i.e., the unbounded perturbation may change the stability type of a fixed point within the persisting slow manifold. For simplicity suppose $\delta = 0$ and consider only real (u, v, w). In this case the reduced equation is a gradient dynamical system with a double well potential, and $u = 1$ is a spatially homogenous equilibrium of the reduced equation. It is attracting since it lies at the minimum of the well. A calculation reveals that the point $(u, v, w) = (1, h^0(1)) = (1, \mu, 1)$ is an equilibrium of the full equations for all $\varepsilon > 0$. Nevertheless, it need not retain the stability type of the $\varepsilon = 0$ limit. Risken and Nummedal [25] showed that the fixed point is unstable for large λ for all positive ε and the number of linearly unstable modes diverges like $1/\varepsilon$. Thus the divergence between the formal limit and the full system is dramatic for small ε.

6.6. Conclusions. We have developed a geometric method of studying the singularly perturbed Maxwell–Bloch equations. The main merit of this method is that it rigorously separates the dynamics of this problem into slow and fast evolution. The geometric principles underlying the method are simple and thus it should be of use in other problems. However, the Maxwell–Bloch equations have several simplifying features and there are often many technical difficulties inherent in a rigorous analysis of PDEs with multiple scales. Thus transporting these ideas to other PDEs will be a difficult (but rewarding) task. Moreover, we have shown that global invariant manifolds with infinite dimension and codimension arise naturally in evolution equations with two scales. One may rigorously find reduced equations for such systems, but these are functional differential equations, and naive approximations to these equations seem to fail. There are several subtle features in geometric singular perturbation theory in infinite dimensions, and the Maxwell–Bloch equations illustrate some of these in a setting with few technicalities.

There have been several recent developments in geometric singular perturbation theory for PDE. We mention the work by Li et al. [21], Haller [17], and Zeng [29] on the damped and driven nonlinear Schrödinger equation. The motivation and methods are different there: in that case the $\varepsilon = 0$ limit is integrable, and a lot of effort is expended in solving problems associated with nonhyperbolicity and weak hyperbolicity. We also mention that Hale, Raugel, Sell, and coworkers have studied PDE in thin domains (see the references in [24]).

Acknowledgments. We thank Professors N. Ercolani, J. Moloney, and G. Kovacic for initiating and encouraging our interest in the Maxwell–Bloch equations. It is also a pleasure to thank Dr. Chongchun Zeng for his assistance. We thank Professor J. Mallet-Paret for pointing out the connection with PDE in thin domains, and Professor E. Titi for bringing to our attention the work by M. Marion. The authors are indebted to the anonymous referees for spotting errors and suggesting several improvements in the manuscript.

REFERENCES

[1] F. T. Arecchi, *Instabilities and chaos in active systems*, in Instabilities and Chaos in Quantum Optics, F. T. Arecchi and R. G. Harrison, eds., Springer-Verlag, Berlin, 1987, pp. 9–48.

[2] J. M. Ball, *Saddle point analysis for an ordinary differential equation in a Banach space and an application to dynamic buckling of a beam*, in Nonlinear Elasticity, J. Dickey, ed., Academic Press, New York, 1974, pp. 93–160.

[3] P. W. Bates and C. K. R. T. Jones, *Invariant manifolds for semilinear partial differential equations*, in Dynamics Reported, Vol. 2, Dynam. Report. Ser. Dynam. Systems Appl. 2, Wiley, Chichester, UK, 1989, pp. 1–38.

[4] P. W. Bates, K. Lu, and C. Zeng, *Existence and Persistence of Invariant Manifolds for Semiflows in Banach Spaces*, Mem. Amer. Math. Soc. 135, AMS, Providence, RI, 1998.

[5] G. Q. Chen, C. D. Levermore, and T.-P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., 47 (1994), pp. 787–830.

[6] S. N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1983.

[7] S. N. Chow, X.-B. Lin, and K. Lu, *Smooth invariant foliations in infinite-dimensional spaces*, J. Differential Equations, 94 (1991), pp. 266–291.

[8] S. N. Chow and K. Lu, *Invariant manifolds for flows in Banach spaces*, J. Differential Equations, 74 (1988), pp. 285–317.

[9] P. Constantin, C. Foias, and J. D. Gibbon, *Finite dimensional attractor for the laser equations*, Nonlinearity, 2 (1989), pp. 241–269.

[10] C. R. Doering, J. Elgin, J. D. Gibbon, and D. D. Holm, *Finite dimensionality in the laser equations in the good cavity limit*, Phys. Lett. A, 129 (1988), pp. 310–316.

[11] N. Fenichel, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J, 21 (1971), pp. 193–226.

[12] N. Fenichel, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.

[13] C. Foias, G. R. Sell, and R. Temam, *Inertial manifolds for nonlinear evolutionary equations*, J. Differential Equations, 73 (1988), pp. 309–353.

[14] J. Hadamard, *Sur l'iteration et les solutions asymptotique des equations differentielles*, Bull. Soc. Math. France, 29 (1901), pp. 224–228.

[15] H. Haken, *Laser Theory*, Springer-Verlag, Berlin, 1983.

[16] H. Haken and H. Ohno, *Theory of ultra-short laser pulses*, Optics Comm., 16 (1976), pp. 205–208.

[17] G. Haller, *Homoclinic jumping in the perturbed nonlinear Schrödinger equation*, Comm. Pure Appl. Math., 52 (1999), pp. 1–47.

[18] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, New York, 1981.

[19] C. K. R. T. Jones, *Geometric singular perturbation theory*, in Dynamical Systems (Montecatini Terme 1994), Lecture Notes in Math. 1609, Springer-Verlag, Berlin, 1995, pp. 44–118.

[20] G. Kovacic, *personal communication.*

[21] Y. Li, D. W. McLaughlin, J. Shatah, and S. Wiggins, *Persistent homoclinic orbits for perturbed nonlinear Schrödinger equations*, Comm. Pure Appl. Math., 49 (1996), pp. 1175–1255.

[22] M. Marion, *Inertial manifolds associated to partly dissipative reaction-diffusion systems*, J. Math. Anal. Appl., 143 (1989), pp. 295–326.

[23] G. Menon, *Gevrey class regularity for the attractor of the laser equations*, Nonlinearity, 12 (1999), pp. 1505–1510.

[24] G. Raugel, *Dynamics of partial differential equations on thin domains*, in Dynamical Systems (Montecatini Terme 1994), Lecture Notes in Math. 1609, Springer-Verlag, Berlin, 1995, pp. 208–315.

[25] H. Risken and J. Nummedal, *Self-pulsing in lasers*, J. Appl. Phys., 39 (1968), pp. 4662–4672.

[26] I. E. Segal, *Nonlinear semigroups*, Ann. Math., 78 (1963), pp. 339–364.

[27] T. Wettergren, *Near Integrable Dynamics of the Maxwell-Bloch Equations*, Ph.D. thesis, Renesselaer Polytechinc Institute, Troy, NY, 1995.

[28] J. Xin and J. Moloney, *Global weak solutions and attractors of the three dimensional Maxwell-Bloch two level laser system*, Comm. Math. Phys., 179 (1996), pp. 511–528.

[29] C. Zeng, *Homoclinic orbits for a perturbed nonlinear Schrödinger equation*, Comm. Pure Appl. Math., 53 (2000), pp. 1222–1283.

SIAM J. Math. Anal.
Vol. 33, No. 2, pp. 347–368

RIDGELETS AND THE REPRESENTATION OF MUTILATED SOBOLEV FUNCTIONS*

EMMANUEL J. CANDES†

Abstract. We show that ridgelets, a system introduced in [E. J. Candes, *Appl. Comput. Harmon. Anal.*, 6 (1999), pp. 197–218], are optimal to represent smooth multivariate functions that may exhibit linear singularities. For instance, let $\{u \cdot x - b > 0\}$ be an arbitrary hyperplane and consider the singular function $f(x) = 1_{\{u \cdot x - b > 0\}} g(x)$, where g is compactly supported with finite Sobolev L_2 norm $\|g\|_{H^s}$, $s > 0$. The ridgelet coefficient sequence of such an object is as sparse as if f were without singularity, allowing optimal partial reconstructions. For instance, the n-term approximation obtained by keeping the terms corresponding to the n largest coefficients in the ridgelet series achieves a rate of approximation of order $n^{-s/d}$; the presence of the singularity does not spoil the quality of the ridgelet approximation. This is unlike all systems currently in use, especially Fourier or wavelet representations.

Key words. Sobolev spaces, Fourier transform, singularities, ridgelets, orthonormal ridgelets, nonlinear approximation, sparsity

AMS subject classifications. 41A46, 42B99

PII. S003614109936364X

1. Introduction.

1.1. Ideal representations of Sobolev classes.

It is well known that trigonometric series and wavelets are well adapted to represent functions taken from L_2 Sobolev classes [1]. For a nonnegative integer s, the L_2 Sobolev norm is

$$\|f\|_{H^s}^2 = \|f\|_2^2 + \|f^{(s)}\|_2^2,$$

where $f^{(s)}$ is the sth derivative of f; and, more generally, the norm of f is defined by means of the Fourier transform; let $\mathcal{F}$ be the classical Fourier transform,

$$(\mathcal{F}f)(\xi) = \hat{f}(\xi) = \int f(x) e^{-ix\cdot\xi}\, dx; \tag{1.1}$$

then,

$$\|f\|_{H^s}^2 = \int |\hat{f}(\xi)|^2 (1 + |\xi|^{2s})\, d\xi$$

when $s > 0$ is not necessarily an integer. (Of course, when s is an integer, the two definitions are equivalent thanks to the Plancherel formula; see [13], for example.)

Both wavelet and Fourier bases provide unconditional bases for these Sobolev spaces H^s defined, say, on the torus. Abstractly, a basis $(\phi_i)_{i\in\mathcal{I}}$ is an unconditional

[*] Received by the editors November 3, 1999; accepted for publication (in revised form) December 16, 2000; published electronically July 19, 2001. This research was supported by National Science Foundation grants DMS 98-72890 (KDI) and DMS 95-05151 and by AFOSR MURI 95-P49620-96-1-0028.

http://www.siam.org/journals/sima/33-2/36364.html

[†] Department of Statistics, Stanford University, Stanford, California 94305-4065 (emmanuel@stat.stanford.edu). Current address: Applied and Computational Mathematics, California Institute of Technology, Mail Code 217-50, Pasadena, CA 91125 (emmanuel@acm.caltech.edu).

basis for a functional class $\mathcal{F}$ if shrinking the coefficients preserves the norm of the object: i.e., if we let

$$\theta_i(f) = \langle f, \phi_i \rangle$$

and consider

$$\tilde{f} = \sum_i \theta_i' \phi_i, \quad |\theta_i'| \le |\theta_i|,$$

then

$$\|\tilde{f}\|_{\mathcal{F}} \le C\, \|f\|_{\mathcal{F}}.$$

We quote Donoho [8]: "An orthogonal basis of L_2 which is also an unconditional basis of a functional space $\mathcal{F}$ is an optimal basis for compressing, estimating, and recovering functions in $\mathcal{F}$."

For instance, suppose that f is a function defined on the circle $\mathbf{T}$ with bounded Sobolev norm and let f_n be the n-term trigonometric nonlinear approximation of f obtained by keeping the terms corresponding to the n largest coefficients in the expansion. Then,

$$\|f - f_n\|_2 \le C\, n^{-s} \|f\|_{H^s(T)}.$$

The same is true for nice periodic wavelets and essentially no orthogonal basis would give a better rate of approximation: that is, for any orthobasis $(\phi_i)_{i \in \mathcal{I}}$, let $Q_n(f)$ be the best n-term approximation in that basis

$$Q_n(f) = \arg\min \|f - g\|_2, \quad g = \sum_{m=1}^{n} \lambda_m \phi_{i_m};$$

then, letting $\mathcal{F}$ be the Sobolev ball $\mathcal{F} = \{f,\ \|f\|_{H^s(T)} \le 1\}$, there is a lower bound on the error of approximation

$$\sup_{f \in \mathcal{F}} \|f - Q_n(f)\|_2 \ge C\, n^{-s}.$$

Another instance of this property is that in any orthobasis $(\phi_i)_{i \in \mathcal{I}}$ the number of terms greater than $1/n$ is greater than $c \cdot n^{2/(2s+1)}$. In both Fourier and wavelet bases, $n^{2/(2s+1)}$ is the order of the number of coefficients that exceed $1/n$, and in this sense we may say that these bases are the most "economical" for representing elements from $H^s(T)$.

1.2. Singularities: The one-dimensional case. However, these nice properties are very fragile. For instance, it is well known that trigonometric series provide poor reconstructions of discontinuous functions. On the interval $[0, 1]$, let f be the periodic function defined by $f(t) = t - H(t - t_0)$, where $H(t)$ is the step function $1_{\{t>0\}}$. The best L_2 n-term approximation of f by trigonometric series gives only an L_2 error of order $O(n^{-1/2})$. This is a general fact: if g is a nice function taken from the Sobolev class H^s (with support contained in (0,1)), then the rate of approximation of $H(t - b)g(t)$ is no better than $O(n^{-1/2})$. The discontinuity spoils the representation, and we need a lot of different terms to reconstruct the discontinuity with good accuracy. (This phenomenon is well known from engineers and is often referred to as the Gibbs phenomenon or ringing effect.)

One of the reasons why wavelets are so attractive is that they are the best bases for representing objects composed with singularities (see the discussion of Mallat's heuristics in [8]). As an example, our simple discontinuous object $H(\cdot - b)g(\cdot)$ has a rate of approximation in a nice wavelet basis of order $O(n^{-s})$. Whereas the singularity had a dramatic effect on the sparsity of Fourier coefficients, it does not affect the sparsity of wavelet coefficients as the number of wavelet coefficients exceeding $1/n$ is still of order $n^{2/(2s+1)}$. The singularity does not spoil the wavelet representation. This miracle may explain the spread of wavelet methods in data compression, statistical estimation, inverse problems, etc., as in practical applications the signals that are to be recovered exhibit these kinds of discontinuities (see the survey paper [11]).

1.3. Singularities: The higher-dimensional case. Under a certain viewpoint, however, the picture changes dramatically when the dimension is greater than one. On $[0,1]^d$, suppose now that we want to represent the simple object

$$f(x) = H(u \cdot x - t_0)g(x), \quad g \in H^s \text{ and supp } g \subset [0,1]^d. \tag{1.2}$$

The object f is singular on the hyperplane $u \cdot x = t_0$ (u is a unit vector) but may be very smooth elsewhere. Then, the number of wavelet coefficients exceeding $1/n$ is greater than $n^{2(1-1/d)}$, yielding L_2 rates of approximation only of order $O(n^{-\frac{1}{2(d-1)}})$. This lower bound holds even when g is as nice as we want, i.e., $g \in C^\infty$. Translated into the framework of image compression, it says that both wavelet bases and Fourier bases are severely inefficient at representing edges in images. Wavelets can deal with point-like phenomena, but they cannot deal with line-like phenomena in dimension 2, plane-like phenomena in dimension 3, etc.

In harmonic analysis, there has recently been much interest in finding new dictionaries and ways of representing functions by linear combinations of elements of those. Examples include wavelets, wavelet-packets, Gabor functions, brushlets, etc. The purpose of this paper is to show that ridgelets, a system introduced by [4], are as efficient for representing objects with discontinuities like (1.2) as wavelets are for representing discontinuous functions in one dimension.

1.4. Achievements and overview. The ridgelet construction will briefly be reviewed in section 2. In a nutshell, a ridgelet is a ridge function of the form

$$\psi_{a,u,b}(x) = \frac{1}{a^{1/2}}\psi\left(\frac{u \cdot x - b}{a}\right), \quad a > 0, u \in S^{d-1}, b \in \mathbb{R}, \tag{1.3}$$

where ψ is univariate and oscillatory. The fundamental result is that there is a discrete family (ψ_{a_n,u_n,b_n}) which is a frame for L_2 spaces of compactly supported functions. (We will simply refer to this family as ψ_n.) The frame property says that for any element $f \in L_2[0,1]^d$ there exist two constants $A, B > 0$ with the property

$$A\,\|f\|^2 \le \sum_n |\langle f, \psi_n\rangle|^2 \le B\,\|f\|^2.$$

A consequence of the previous display is the existence of a dual set of ridgelets $(\widetilde{\psi_n})$ (the dual frame) and of the decomposition

$$f = \sum_n \langle f, \widetilde{\psi_n}\rangle \psi_n = \sum_n \langle f, \psi_n\rangle, \widetilde{\psi_n} \tag{1.4}$$

with equality holding in an L_2 sense.

To measure the sparsity of an arbitrary sequence (θ_n), we will use the weak-ℓ_p or Marcinkiewicz quasi norm, defined as follows: let $|\theta|_{(n)}$ be the nth largest entry in the sequence $(|\theta_n|)$; we set

$$|\theta|_{w\ell_p} = \sup_{n>0} n^{1/p} |\theta|_{(n)}. \tag{1.5}$$

Equipped with a nice ridgelet frame, the key result of our paper (section 4) is the following: let us consider a template f such as in (1.2) and let α ($\alpha_n = \langle f, \psi_n \rangle$) denote the ridgelet coefficient sequence of f. Then, the sequence α is sparse as if f were not singular in the sense that

$$\|\alpha\|_{w\ell_p} \le C \, \|g\|_{H^s} \quad \text{with } 1/p = s/d + 1/2, \tag{1.6}$$

where the constant C does not depend on f; or equivalently, the number of ridgelet coefficients exceeding $1/n$ is bounded by $C\, n^p \, \|g\|_{H^s}$. (Throughout the paper, the letter C will denote a positive constant whose value may differ at different occurrences, even within a single formula.) There might be some ambiguity about the notation $\|g\|_{H^s}$ since g is not uniquely determined by f. In this paper, we will implicitly take the norm $\|g\|_{H^s}$ as being the minimum norm of all those elements in H^s whose restriction to $\{u \cdot x > t_0\}$ coincide with f; i.e.,

$$\|g\|_{H^s} := \inf\{\|h\|_{H^s}, \, f(x) = H(u \cdot x - t_0) h(x), \, \operatorname{supp} h \subset [0,1]^d\}.$$

There is a direct consequence of this result. Consider the n-term f_n ridgelet approximation obtained by extracting from the exact series (1.4) the terms corresponding to the n largest coefficients. Then,

$$\|f - f_n\| \le C \, n^{-s/d} \, \|g\|_{H^s}, \tag{1.7}$$

where, again, the constant C is independent of f. *The presence of the singularity does not ruin the sparsity of the ridgelet series.* This is unlike wavelet or Fourier analysis. Hence, we have a very concrete, constructive, and stable procedure—namely, the thresholding of ridgelet coefficients—to obtain near-optimal nonlinear approximations. The author is not aware of any other system with similar features.

In dimension 2, Donoho introduced an orthonormal basis, closely related to the ridgelet system, that he calls "orthonormal ridgelets." Section 5 will show that both results (1.6) and (1.7) continue to hold with orthonormal ridgelets in place of "pure" ridgelets.

1.5. Methodology. The method that is used to prove (1.6) and (1.7) involves the study of the Fourier transform along rays going through the origin (section 3). Before we proceed further, (r, θ) will index the standard polar coordinates system and throughout the paper we will abuse notation in writing $f(r, \theta)$ instead of $(f \circ \mathcal{C})(r, \theta)$, where $\mathcal{C}$ is the change of coordinates from polar to cartesian. In two dimensions, let us now consider the singular function f defined by

$$f(x_1, x_2) = 1_{\{x_1 > 0\}} \, g(x_1, x_2)$$

with g in H^s, $s \in \mathbb{N}$, and supp $g \subset [0,1]^d$. The argument relies on a bound that is available on the integral over the "polar" segment $\{(r, \theta), \, 2^j \le r \le 2^{j+1}\}$ of the squared modulus of the Fourier transform. Indeed, there exists a constant C not depending on f such that

$$\int_{2^j \le r \le 2^{j+1}} |\hat{f}(r,\theta)|^2 \, dr \\ \le C\, \epsilon_j^2(\theta) 2^{-j} 2^{-2js} \|g\|_{H^s}^2 + C\, 2^{-j} \min(1, 2^{-2js} |\sin\theta|^{-2s}) \|g\|_{H^s}^2 \tag{1.8}$$

with $\sum_j \int_0^{2\pi} \epsilon_j^2(\theta)\, d\theta \le 1$. A d-dimensional version of (1.8) will be given in section 3.

The singularity $1_{\{x_1 > 0\}}$ causes the Fourier transform to decay very slowly in the critical directions $\theta = 0, \pi$. (This set of directions is sometimes referred to as the wavefront.) Indeed, for $\theta = 0$, say, $|\hat{f}(r,\theta)| \sim r^{-1}$ and, therefore, for this critical value of θ, $\int_{2^j \le r \le 2^{j+1}} |\hat{f}(r,\theta)|^2 \, dr \sim 2^{-j}$, which is the content of (1.8). However, this effect is really local and our estimate (1.8) pictures the decay of the Fourier transform as θ moves away from the singular rays. The result is nonasymptotic since it describes the situation at a finite distance 2^j ($j \ge 0$) from the origin. For instance, in dimension 2 the order of magnitude of the modulus of the Fourier transform at a point with polar coordinates $(2^j, \theta)$ is $2^{-j(s+1)} |\sin\theta|^{-s}$. It is interesting to observe that the smoothness of the object governs the size of the Fourier transform as θ approaches $0, \pi$. Although this phenomenon may not have been extensively studied in the literature, it perhaps corresponds to some new kind of microlocal analysis and we believe that this is of independent interest.

The localization of the Fourier transform near the wavefront is the key property driving our main results (1.6) and (1.7). Extensions and limitations of these results will be discussed in section 6.

2. Ridgelets. In this section, $\hat{g}$ will denote the Fourier transform of g. In d dimensions, the ridgelet construction starts with a univariate function ψ satisfying an oscillatory condition, namely,

$$\int |\hat{\psi}(\xi)|^2 / |\xi|^d \, d\xi < \infty. \tag{2.1}$$

A ridgelet is a function of the form

$$\frac{1}{a^{1/2}} \psi\left(\frac{u \cdot x - b}{a}\right), \tag{2.2}$$

where a and b are scalar parameters and u is a vector of unit length. In what follows, we will suppose that ψ is normalized so that $\int |\hat{\psi}(\xi)|^2 |\xi|^{-d} d\xi = 1$. Of course, a ridgelet is a ridge function whose profile displays an oscillatory behavior (like a wavelet). A ridgelet has a scale a, an orientation u, and a location parameter b. Ridgelets are concentrated around hyperplanes: roughly speaking, the ridgelet (2.2) is supported near the strip $\{x, |u \cdot x - b| \le a\}$.

Remarkably, one can represent any function as a superposition of these ridgelets. Define the ridgelet coefficients

$$\mathcal{R}_f(a,u,b) = \int f(x)\, a^{-1/2} \psi\left(\frac{u \cdot x - b}{a}\right) dx; \tag{2.3}$$

then, for any $f \in L_1 \cap L_2(\mathbb{R}^d)$, we have

$$f(x) = (2\pi)^{-(d-1)} \int \mathcal{R}_f(a,u,b) a^{-1/2} \psi\left(\frac{u \cdot x - b}{a}\right) d\mu(a,u,b), \tag{2.4}$$

where $d\mu(a,u,b) = da/a^{d+1}\,du\,db$ (du being the uniform measure on the sphere). Furthermore, this formula is stable as one has a Parseval relation

$$\|f\|_2^2 = (2\pi)^{-(d-1)} \int |\mathcal{R}_f(a,u,b)|^2 d\mu(a,u,b). \tag{2.5}$$

Similar to the continuous transform, there is a discrete transform. Consider the following discrete collection of ridgelets:

$$\{\psi_{j,\ell,k}(x) = 2^{j/2}\psi(2^j u_{j,\ell}\cdot x - kb_0),\ j \geq j_0, u_{j,\ell} \in \Sigma_j, k \in \mathbb{Z}\}. \tag{2.6}$$

The scale a and location parameter b are discretized dyadically, as in the theory of wavelets. However, unlike wavelets, ridgelets are directional and here the interesting aspect is the discretization of the directional variable u. This variable is sampled at increasing resolution, so that at scale j the discretized set Σ_j is a net of nearly equispaced points at a distance of order 2^{-j}. A detailed exposition on the ridgelet construction is given in [4]. In two dimensions, for instance, a ridgelet is of the form

$$\{\, 2^{j/2}\psi(2^j(x_1\cos\theta_{j,\ell} + x_2\sin\theta_{j,\ell} - 2\pi k 2^{-j}))\,\}_{(j\geq j_0,\ell,k)},$$

where the directional parameter $\theta_{j,\ell}$ is sampled with increasing angular resolution at increasingly fine scales, something like the following:

$$\theta_{j,\ell} = 2\pi\ell 2^{-j}.$$

The key result [4] is that the discrete collection $(\psi_{j,\ell,k})$ is a frame for square integrable functions supported on the unit cube. There exist two constants A and B such that for any $f \in L_2([0,1]^d)$, we have

$$A\,\|f\|_{L_2}^2 \leq \sum_{j,\ell,k} |\langle f, \psi_{j,\ell,k}\rangle|^2 \leq B\,\|f\|_{L_2}^2. \tag{2.7}$$

The previous equation says that the datum of the ridgelet transform at the points $(a,u,b) = (2^j, u_{j,\ell}, k2^{-j})$—with the parameter range as in (2.6)—suffices to reconstruct the function perfectly. In this sense, this is analogous to the Shannon sampling theorem for the reconstruction of bandlimited functions. Indeed, standard arguments show that there exists a dual collection $(\tilde{\psi}_{j,\ell,k})$ with the property

$$f = \sum_{j,\ell,k} \langle f, \tilde{\psi}_{j,\ell,k}\rangle \psi_{j,\ell,k} = \sum_{j,\ell,k} \langle f, \psi_{j,\ell,k}\rangle \tilde{\psi}_{j,\ell,k}, \tag{2.8}$$

where the notation $\langle\cdot,\cdot\rangle$ stands here and throughout the remainder of this paper for the usual inner product of L_2: $\langle f, g\rangle = \int f(x)g(x)dx$.

At times, we will use the compact notation ψ_ν ($\nu \in \mathcal{N}$) for our ridgelet frames and, therefore, we will keep in mind that the index runs ν through an enumeration of the triples (j,ℓ,k).

3. Localization of the Fourier transform. The purpose of this section is to quantify the size of the Fourier transform of an object f, where f is given by

$$f(x) = H(x_1)\,g(x),$$

where g is compactly supported and with finite Sobolev norm (recall $H(t) = 1_{\{t>0\}}$).

To formulate our statement in d dimensions, we introduce the spherical coordinates defined by $x_1 = r\cos\theta_1$, $x_2 = r\sin\theta_1\cos\theta_2$, $\dots$, $x_d = r\sin\theta_1\sin\theta_2\dots\sin\theta_{d-1}$, $0 \le \theta_1,\dots,\theta_{d-2} \le \pi$, $0 \le \theta_{d-1} < 2\pi$. In what follows, we will simply refer to $(\theta_2,\dots,\theta_{d-1})$ as φ, and $d\varphi$ will denote the element of the surface area on S^{d-2}, i.e., $d\varphi = \sin\theta_2^{d-3}\dots\sin\theta_{d-2}d\theta_2\dots d\theta_{d-1}$. With these notations, the uniform measure du on the sphere may thus be rewritten as $du = (\sin\theta_1)^{d-2}\,d\theta_1 d\varphi$. From now on, we will often refer to a unit vector u by means of its polar coordinates (θ,φ), $\theta\in[0,\pi]$, $\varphi\in S^{d-2}$.

We now state our d-dimensional localization result about the modulus of the Fourier transform.

Theorem 3.1. *Let f be given by $f(x) = H(x_1)\,g(x)$ with g in H^s, $s = 0,1,2,\dots$, and* supp $g \subset [-1,1]^d$, *and put $\sigma = s + (d-2)/2$. Then, there exists a universal constant C such that for any $j \ge 0$,*

$$\text{(3.1)}\quad \int_{2^j\le r\le 2^{j+1}}\int |\hat f(r,\theta,\varphi)|^2\,dr d\varphi \\ \le C\,\epsilon_j^2(\theta)2^{-j}2^{-2j\sigma}\|g\|_{H^s}^2 + C\,2^{-j}\min(1,2^{-2j\sigma}|\sin\theta|^{-2\sigma})\|g\|_{H^s}^2,$$

where $\sum_j |S^{d-2}|\int\epsilon_j^2(\theta)(\sin\theta)^{d-2}d\theta \le 1$.

As we emphasized earlier, the Fourier transform decays very slowly in the directions $\theta = 0,\pi$ because of the singularity H. However, (3.1) is not a statement about the decay of $\hat f$ along the singular rays $\theta = 0,\pi$; rather it is about the decay of the Fourier transform as θ moves away from the critical directions $\theta = 0,\pi$. Roughly speaking, the order of magnitude of the modulus of the Fourier transform at a point with polar coordinates $(2^j,\theta)$ is $2^{-j(\sigma+1)}|\sin\theta|^{-\sigma}$ with $\sigma = s + (d-2)/2$.

Remark. The inequality involves a regular term (the first term of the right-hand side of (3.1)) as if one were simply analyzing an object from H^s and a singular term (the second one) essentially due to the discontinuity across the hyperplane $x_1 = 0$.

Proof. We will prove the result by induction. The result is true for $s = 0$ since letting $I_j(\theta)$ be the left-hand side of (3.1)

$$I_j(\theta) \equiv \int_{2^j\le r\le 2^{j+1}}\int |\hat f(r,\theta,\varphi)|^2\,dr d\varphi,$$

we have, by definition,

$$\sum_{j\ge0}2^{j(d-1)}\int I_j(\theta)(\sin\theta)^{d-2}\,d\theta = \sum_{j\ge0}2^{j(d-1)}\int_{2^j}^{2^{j+1}}\int|\hat f(r,\theta,\varphi)|^2\,dr d\theta d\varphi \\ \le \sum_{j\ge0}\int_{2^j\le|\xi|\le 2^{j+1}}|\hat f(\xi)|^2\,d\xi \le \|f\|_{L_2}^2 \le \|g\|_{L_2}^2.$$

Assume now that the result holds until $n-1$ ($n\in\mathbb{N}$), and take $g\in H^n$. For any tempered distribution in $\mathbb{R}^d$ S, we have the well-known relationship

$$\mathcal{F}\{\partial_\ell S\} = i\xi_\ell\hat S,$$

where in the previous display $i^2 = -1$, and ∂_ℓ is the partial derivative with respect to the ℓth coordinate. We will simply apply this formula to the tempered distribution $f = H\,g$. First, for any $1\le\ell\le d$, we have

$$\text{(3.2)}\qquad \partial_\ell f = H\,\partial_\ell g + g\,\partial_\ell H.$$

We observe that the second term, $g\,\partial_\ell H$, is nonzero only if $\ell = 1$ in which case it is a distribution supported on $x_1 = 0$, namely, $g\,\delta_{\{x_1=0\}}$. Let h be the restriction of g on $x_1 = 0$. By the trace theorem [15] we know that h is in $H^{n-1/2}(\mathbb{R}^{d-1})$ and, more precisely,

$$\|h\|_{H^{n-1/2}} \le C\,\|g\|_{H^n}.$$

Let us now choose $u = \xi/|\xi|$ and let $\xi = (\xi_1, \xi')$ so that $\xi' = \pi(\xi)$, where π is the orthogonal projection onto $\xi_1 = 0$. For this particular choice of u, we have

$$i|\xi|\hat{f}(\xi) = u\cdot\mathcal{F}\{\nabla f\}(\xi) = u\cdot\mathcal{F}\{H\,\nabla g\}(\xi) + \xi_1/|\xi|\,\hat{h}(\pi(\xi)) \tag{3.3}$$

since the Fourier transform of $g\,\delta_{\{x_1=0\}}$ is given by $\hat{h}(\pi(\xi)) = (\hat{h}\circ\pi)(\xi)$. The first term of the right-hand side of (3.3) is effortlessly going through the induction step. Indeed, we have

$$|u\cdot\mathcal{F}\{H\,\nabla g\}|^2(\xi) \le \sum_{i=1}^{d} |\mathcal{F}\{H\,\partial_\ell g\}|^2(\xi);$$

it is clear that for any ℓ, $\partial_\ell g \in H^{n-1}$ and therefore the induction hypothesis implies that

$$\begin{aligned}\int_{2^j\le r\le 2^{j+1}} \int |u\cdot\mathcal{F}\{H\,\nabla g\}|^2(r,\theta,\varphi)\,drd\varphi \\ \le C\,2^{-j}\epsilon_j^2(\theta)2^{-2j(\sigma-1)} + C\,2^{-j}\min(1, 2^{-2j(\sigma-1)}|\sin\theta|^{-2(\sigma-1)}).\end{aligned} \tag{3.4}$$

We split the analysis of the second term of the right-hand side of (3.3) into two separate cases: namely, $\sin\theta \ge 2^{-j}$ and $\sin\theta < 2^{-j}$. In the former case, we have

$$\begin{aligned}\int_{2^j}^{2^{j+1}} \int |(\hat{h}\circ\pi)(r,\theta,\varphi)|^2\,drd\varphi &= \int_{2^j}^{2^{j+1}} \int |\hat{h}(r\sin\theta,\varphi)|^2\,drd\varphi \\ &= |\sin\theta|^{-1} \int_{2^j|\sin\theta|}^{2^{j+1}|\sin\theta|} \int |\hat{h}(\rho,\varphi)|^2\,d\rho d\varphi \\ &\le 2^{-j(d-2)}|\sin\theta|^{-(d-1)} \int_{2^j\le|\xi'|/|\sin\theta|\le 2^{j+1}} |\hat{h}(\xi')|^2\,d\xi'.\end{aligned}$$

The degree of smoothness of h ($h \in H^{n-1/2}$) now allows us to bound the right-hand side of the previous display; i.e.,

$$\sum_{j=-\infty}^{\infty} |2^j\sin\theta|^{2(n-1/2)} \int_{2^j|\sin\theta|\le|\xi'|\le 2^{j+1}|\sin\theta|} |\hat{h}(\xi')|^2\,d\xi' \sim \|h\|^2_{\dot{H}^{n-1/2}} \le C\,\|g\|^2_{H^n},$$

which implies

$$\int_{2^j|\sin\theta|\le|\xi'|\le 2^{j+1}|\sin\theta|} |\hat{h}(\xi')|^2\,d\xi' \le C\,\eta_j^2(\theta)\,|2^j\sin\theta|^{-2(n-1/2)}\,\|g\|^2_{H^n}$$

with $\sum_j \eta_j^2(\theta) \le 1$.

To summarize, we have

$$\int_{2^j \le r \le 2^{j+1}} \int |(\hat{h} \circ \pi)(r, \theta, \varphi)|^2 \, dr d\varphi \le C \, 2^{-2j(\sigma - 1/2)} |\sin \theta|^{-2\sigma} \, \|g\|_{H^s}^2 \tag{3.5}$$

in any dimension $d \ge 2$.

To finish the proof, we simply recall (3.3) which gives the inequality

$$|\hat{f}(\xi)|^2 = 2|\xi|^{-2} \left(|u \cdot \mathcal{F}\{H \, \nabla g\}(\xi)|^2 + |\hat{h}(\pi(\xi))|^2 \right).$$

The polar integral of each term of the right-hand side of this inequality is bounded via (3.4) and (3.5), respectively, yielding the desired conclusion. The case $\sin \theta \ge 2^{-j}$ is now fully proved.

We finally treat the case $\sin \theta < 2^{-j}$. On one hand h is bounded in $H^{n-1/2}$ and therefore in L_2, since $n \ge 1$. On the other hand, h is compactly supported and hence

$$\sup_{|\xi'| \le 1} |\hat{h}(\xi')| \le \|h\|_{L_1} \le C \, \|h\|_{L_2} \le C \, \|g\|_{H^n}.$$

In this case, we simply write

$$\begin{aligned} \int_{2^j \le r \le 2^{j+1}} \int |\hat{h}(r \sin \theta, \varphi)|^2 \, dr d\varphi &\le 2^j |S^{d-2}| \sup_{2^j |\sin \theta| \le |\xi'| \le 2^{j+1} |\sin \theta|} |\hat{h}(\xi')|^2 \\ &\le C \, 2^j \|g\|_{H^n}^2, \end{aligned}$$

and the result for $\sin \theta < 2^{-j}$ now follows from (3.3). The proof of the theorem is complete. □

4. Main result. In this section, we will suppose that we are given a ridgelet frame satisfying the following mild assumptions.

1. ψ is R times differentiable and has vanishing moments through order D; $\min(R, D) \ge s + (d-1)/2$.

2. ψ is of rapid decay; namely, for any $\gamma > 0$ and $0 \le r \le R$, one can find a constant C such that

$$|\psi^{(r)}(t)| \le C \cdot (1 + |t|)^{-\gamma}.$$

The sequence of ridgelet coefficients of a given function f will be denoted by α: $\alpha_{j,\ell,k} = \langle f, \psi_{j,\ell,k} \rangle$.

We state our main result.

THEOREM 4.1. *Let $g \in H^s$, $s > 0$, with* supp $g \subset [-1,1]^d$ *and put $f(x) = H(u \cdot x - b) \, g(x)$, where H is the step function $H(t) = 1_{\{t > 0\}}$. Then, the ridgelet coefficient sequence α of f satisfies*

$$\|\alpha\|_{w\ell_{p^*}} \le C \, \|g\|_{H^s} \quad \textit{with} \quad 1/p^* = s/d + 1/2,$$

where d is the dimension of the space.

Preliminary remark. For any (j, ℓ, k), we have the following basic inequality:

$$|\alpha_{j,\ell,k}| \le 2^{j/2} (1 + |k|)^{-\gamma} \|f\|_2, \quad |k| \ge 2^{j+1},$$

because of the rapid decay of ψ. Indeed, we have

$$|\psi_{j,\ell,k}(x)| \le C \, (1 + 2^j |u_{j,\ell} \cdot x - k 2^{-j}|)^{-\gamma},$$

and, therefore, it is not hard to check that for $|k| \geq 2^{j+1}$

$$\sup_{[-1,1]^d} |\psi_{j,\ell,k}(x)| \leq C\, 2^{j/2}(1+|k|)^{-\gamma}.$$

Our claim is then a simple consequence of this last inequality. Thus, if ψ has a sufficient decay, then the subsequence $\{(\alpha_{j,\ell,k}),\ k \geq 2^{j+1}\}$ is in ℓ_p for any $p > 0$; hence it is enough to restrict our attention to the set $|k| \leq 2^{j+1}$.

In order to prove the theorem, we will need a result which is a corollary of Theorem 3.1.

Corollary 4.2. *Under the assumptions of Theorem* 3.1*, the ridgelet coefficient sequence α of f may be decomposed as*

$$\alpha_{j,\ell,k} = a_{j,\ell,k} + b_{j,\ell,k},$$

where the sequences a and b enjoy the following properties.

1. *The sequence a verifies*

$$\sum_{\ell,k} |a_{j,\ell,k}|^2 \leq C\, \epsilon_j^2 2^{-2js}\, \|g\|_{H^s}^2 \tag{4.1}$$

with $\sum_j \epsilon_j^2 \leq 1$, and

2. *the sequence b is localized both in angle and in location.*

(i) *Localization in angle. For $1 \leq m < j$, let $\Lambda_{j,m}$ be the set of indices such that*

$$\Lambda_{j,m} := \{\ell,\ 2^{-m} \leq |\sin\theta_{j,\ell}| \leq 2^{-m+1}\} \tag{4.2}$$

(for $m = j$, we will take $\Lambda_{j,m}$ to be $\{\ell,\ |\sin\theta_{j,\ell}| \leq 2^{-(j-1)}\}$); then,

$$\sum_{\ell\in\Lambda_{j,m}} \sum_k |b_{j,\ell,k}|^2 \leq C\, 2^{-j}\, 2^{-(j-m)(2s-1)}\, \|g\|_{H^s}^2. \tag{4.3}$$

(ii) *Localization in ridge location. For any $n > 0$, there is a constant C (independent of f) such that*

$$|b_{j,\ell,k}| \leq C\, 2^{j/2}\left(1 + \big||k| - |2^j \sin\theta_{j,\ell}|\big|\right)^{-n}\, \|g\|_{H^s}. \tag{4.4}$$

Not surprisingly, this decomposition involves a regular and a singular contribution as well.

Proof of Corollary 4.2. Again, we prove the result by induction. For any compactly supported element of L_2, we have

$$\sum_j \sum_{\ell,k} |\alpha_{j,\ell,k}|^2 \leq C\, \|f\|_{L_2}^2 \leq C\, \|g\|_{L_2}^2,$$

which proves the claim in this case since one can simply take $b \equiv 0$.

Suppose now that the claim is true up to $s - 1 \in \mathbb{N}$ and take g in H^s. Recall that the ridgelet $\psi_{j,\ell,k}$ is given by $2^{j/2}\psi(2^j u_{j,\ell}\cdot x - k)$. The starting point is to express the ridgelet coefficient $\alpha_{j,\ell,k}$ as a line integral in the Fourier domain [4]

$$\alpha_{j,\ell,k} = \int_{\mathbb{R}} \hat{f}(\lambda, u_{j,\ell}) 2^{-j/2}\hat{\psi}(2^{-j}\lambda)e^{-ik2^{-j}\lambda}\, d\lambda, \tag{4.5}$$

where $\hat{f}(\lambda, u) = \hat{f}(\lambda u_1, \dots, \lambda u_d)$. In the previous equation, the range of λ is the real line and not only the positive axis (polar coordinates). However, we can convert (λ, u) to classical polar coordinates (r, θ, φ) via the obvious relationship $(\lambda, u) = (-\lambda, -u)$. The decomposition (3.3) then suggests rewriting $\alpha_{j,\ell,k}$ as

$$\alpha_{j,\ell,k} = a^{(0)}_{j,\ell,k} + b^{(0)}_{j,\ell,k},$$

where

$$a^{(0)}_{j,\ell,k} = 2^{-j}\, u_{j,\ell} \cdot \int_{\mathbb{R}} \mathcal{F}\{H\,\nabla g\}(\lambda, u_{j,\ell}) 2^{-j/2} \frac{\hat{\psi}(2^{-j}\lambda)}{2^{-j}\lambda} e^{-ik2^{-j}\lambda}\, d\lambda$$

and

$$b^{(0)}_{j,\ell,k} = 2^{-j} \cos\theta_{j,\ell} \int_{\mathbb{R}} \hat{h}(\lambda \sin\theta_{j,\ell}, \varphi_{j,\ell}) \frac{\hat{\psi}(2^{-j}\lambda)}{2^{-j}\lambda} e^{-ik2^{-j}\lambda}\, d\lambda.$$

Let Ψ be the primitive of ψ defined by $\Psi(x) = \int_{-\infty}^{x} \psi(t)\, dt$. Then, Ψ satisfies the conditions listed at the beginning of the section (with the obvious modification $\min(R, D) \geq s - 1 + (d-1)/2$) and $\hat{\Psi}(\lambda) = -i\hat{\psi}(\lambda)/\lambda$. Therefore, we may apply the induction hypothesis to the sequence a and obtain

$$a^{(0)}_{j,\ell,k} = 2^{-j} a^{(1)}_{j,\ell,k} + 2^{-j} b^{(1)}_{j,\ell,k},$$

where $a^{(1)}$ and $b^{(1)}$, respectively, satisfy properties (4.1) and (4.3)–(4.4) with $(s-1)$ in place of s. Now, define the sequences a and b by

$$a_{j,\ell,k} = 2^{-j} a^{(1)}_{j,\ell,k}$$

and

$$b_{j,\ell,k} = 2^{-j} b^{(1)}_{j,\ell,k} + b^{(0)}_{j,\ell,k}.$$

It is clear that $a_{j,\ell,k}$ and $2^{-j} b^{(1)}_{j,\ell,k}$ satisfy conditions (4.1) and (4.3)–(4.4), respectively. Thus we need only to check that the sequence $b^{(0)}$ verifies (4.3) and (4.4). In the original domain, $b^{(0)}_{j,\ell,k}$ is given by

$$b^{(0)}_{j,\ell,k} = \langle g\, \delta_{\{x_1=0\}}, \Psi_{j,\ell,k} \rangle$$

and, therefore, with the the same notations as in section 3, i.e., $h(x') = g(0, x')$,

$$|b^{(0)}_{j,\ell,k}| \leq \|h\|_{L_1} \sup_{x \in \text{ supp } g\delta_{\{x_1=0\}}} |\Psi_{j,\ell,k}(x)|.$$

First, it is easy to see that $\Psi_{j,\ell,k}$ is bounded by $C\, 2^{j/2} \left(1 + \big||k| - |2^j \sin\theta_{j,\ell}|\big|\right)^{-n}$ on the support of $g\, \delta_{\{x_1=0\}}$ and second, we have $\|h\|_{L_1} \leq C\|h\|_{L_2} \leq C\|g\|_{H^{1/2}}$ which is bounded since $g \in H^s$, $s \geq 1$. This finishes the verification of (4.4). It remains to check (4.3).

Sampling results. In a separate paper, we have established the following sampling results: let $\alpha_{j,\ell,k}$ be the ridgelet coefficients of a compactly supported distribution S; first,

$$\sum_k |\alpha_{j,\ell,k}|^2 \leq C \int_{\mathbb{R}} |\hat{S}(\lambda, u_{j,\ell})|^2 |\hat{\psi}(2^{-j}\lambda)|^2 (1 + |2^{-j}\lambda|^2)\, d\lambda; \tag{4.6}$$

second, we recall that at scale j, the set of discrete angular variables $\{u_{j,\ell}, \ell \in \Lambda_j\}$ consists of points approximately uniformly distributed on the sphere; for any subset Λ_j' of Λ_j, we have

$$(4.7) \quad \sum_{\ell \in \Lambda_j'} \sum_k |\alpha_{j,\ell,k}|^2 \\ \leq C\, 2^{j(d-1)} \int_{\mathbb{R}} |\hat{\psi}(2^{-j}\lambda)|^2 (1 + |2^{-j}\lambda|^{2d})\, d\lambda \int_{\Sigma_j'} \sum_{|\alpha| \leq d-1} |D^\alpha \hat{S}(\lambda, u)|^2\, du,$$

where Σ_j' is the set of points on the sphere defined by

$$\Sigma_j' \equiv \left\{ u \in S^{d-1}, \inf_{\ell \in \Lambda_j'} \|u - u_{j,\ell}\|_2 \leq 2^{-j} \right\}.$$

Here α is a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d)$ and D^α stands for the classical partial derivative with respect to the cartesian coordinate system $D^\alpha S = \partial_1^{\alpha_1} \ldots \partial_d^{\alpha_d}$. Thus, (4.7) is a kind of uniform sampling inequality. In a nutshell, (4.7) holds because the points $\{u_{j,\ell}, \ell \in \Lambda_j\}$ are quasi-uniformly distributed on the sphere (at a distance of order 2^{-j}); that is, for any point $u \in S^{d-1}$,

$$\#\{\ell, \|u_{j,\ell} - u\|_2 \leq \delta\} \leq C\, 2^{j(d-1)} \delta^{d-1}.$$

We apply this result to the distribution $S = g\, \delta_{\{x_1 = 0\}}$, that is, to the restriction of f to the hyperplane $\{x_1 = 0\}$ (see section 3 for details). The Fourier transform of S is the function $\hat{S} = \hat{h} \circ \pi$ that we introduced in section 3. With $\Lambda_{j,m}$, $0 \leq m < j$, as in (4.2), we have

$$\inf_{\ell \in \Lambda_{j,m}} \|u - u_{j,\ell}\|_2 \leq 2^{-j} \quad \Rightarrow \quad 2^{-m} - 2^{-j} \leq \sin\theta \leq 2^{-m+1} + 2^{-j}$$

and we omit the proof of this simple inclusion. Therefore, in this context (4.7) gives

$$(4.8) \quad \sum_{\ell \in \Lambda_{j,m}} \sum_k |b_{j,\ell,k}^{(0)}|^2 \leq C\, 2^{j(d-1)} \int_{2^{-m} - 2^{-j} \leq \sin\theta \leq 2^{-m+1} + 2^{-j}} I(\theta)\, (\sin\theta)^{d-2}\, d\theta,$$

where $I(\theta)$ is given by

$$\int_{\mathbf{S}^{d-2}} \int_{\mathbb{R}} \sum_{|\alpha| \leq d-1} |D^\alpha \hat{S}(\lambda, \theta, \varphi)|^2 |\hat{\psi}(2^{-j}\lambda)|^2 (1 + |2^{-j}\lambda|^{2d})\, d\lambda d\varphi.$$

Now, if ψ has r vanishing moments and is of regularity r, we have

$$(4.9) \quad \sup_{2^\ell \leq |\lambda| \leq 2^{\ell+1}} |\hat{\psi}(2^{-j}\lambda)| \leq C\, 2^{-|j-\ell| r}.$$

It is then easy to check that

$$(4.10) \quad I(\theta) \leq C\, 2^{-j} 2^{-2j\sigma} |\sin\theta|^{-2\sigma}\, \|g\|_{H^s}^2.$$

To see why this is true, we simply write

$$I(\theta) \leq \sum_\ell \sup_{2^\ell \leq |\lambda| \leq 2^{\ell+1}} |\hat{\psi}(2^{-j}\lambda)|^2 (1 + |2^{-j}\lambda|^{2d}) I_\ell(\theta),$$

where

$$I_\ell(\theta) = \int_{2^\ell \le |\lambda| \le 2^{\ell+1}} \int \sum_{|\alpha| \le d-1} |D^\alpha \hat{S}(\lambda, \theta, \varphi)|^2 \, d\lambda d\varphi.$$

In the proof of Theorem 3.1 (3.5), we obtained

$$\int_{2^\ell \le |\lambda| \le 2^{\ell+1}} \int |\hat{S}(\lambda, \theta, \varphi)|^2 \, d\lambda d\varphi \le C \, 2^\ell 2^{-2\ell\sigma} |\sin\theta|^{-2\sigma} \|g\|_{H^s}^2. \tag{4.11}$$

Now, $D^\alpha \hat{S}$ is the Fourier transform of the distribution $(-i)^{|\alpha|} \, x^\alpha \, S$, which is the restriction of $(-i)^{|\alpha|} \, x^\alpha \, g$ to the hyperplane $\{x_1 = 0\}$. Because g is compactly supported, we have that

$$\|x^\alpha \, g\|_{H^s} \le C \, \|g\|_{H^s}$$

since the multiplication by a C_0^∞ function is a bounded operation from H^s onto itself. Therefore, inequality (4.11) applies to $D^\alpha \hat{S}$, and we have the upper bound

$$I_\ell(\theta) \le C \, 2^\ell 2^{-2\ell\sigma} |\sin\theta|^{-2\sigma} \|g\|_{H^s}^2.$$

Inequality (4.10) comes from the previous inequality together with the size estimate (4.9).

Combining (4.10) and (4.8) finally gives (recall $2\sigma = 2s + d - 2$)

$$\sum_{\ell \in \Lambda_{j,m}} \sum_k |b_{j,\ell,k}^{(0)}|^2 \le C \, 2^{-2js} \, \|g\|_{H^s}^2 \int_{2^{-m} - 2^{-j} \le \sin\theta \le 2^{-m+1} + 2^{-j}} |\sin\theta|^{-2s} \, d\theta,$$

which, in turn, gives the desired conclusion

$$\sum_{\ell \in \Lambda_{j,m}} \sum_k |b_{j,\ell,k}^{(0)}|^2 \le C \, 2^{-m} 2^{-2(j-m)s} \, \|g\|_{H^s}^2.$$

The corollary is established. □

Proof of Theorem 4.1. Let s be a positive integer. Following on Corollary 4.2, to prove that α is in $w_{\ell_{p^*}}$ $(1/p^* = s/d + 1/2)$, it is sufficient to prove that both a and b are in $w_{\ell_{p^*}}$. The membership of a to $w_{\ell_{p^*}}$ follows from well-known arguments and is straightforward.

The $w\ell_{p^*}$ boundedness of the sequence $(b_{j,\ell,k})$ will be deduced from Corollary 4.2. We identify two subsequences corresponding, respectively, to the indices $|k| > 2^{j+1} |\sin\theta_{j,\ell}|$ and $|k| \le 2^{j+1} |\sin\theta_{j,\ell}|$; the interesting contribution concerns the latter subsequence. We prove that

1. the subsequence $\{b_{j,\ell,k}, \, |k| \le 2^{j+1} |\sin\theta_{j,\ell}|\}$ has a finite $w_{\ell_{p^*}}$ norm, and
2. the ℓ_p norm of the subsequence $\{b_{j,\ell,k}, \, |k| > 2^{j+1} |\sin\theta_{j,\ell}|\}$ is bounded for any $p > 0$.

We prove the first assertion. Letting $N(\epsilon)$ be the cardinality of those elements whose absolute value exceeds ϵ, namely,

$$N(\epsilon) = \# \, \epsilon \, \{(j, \ell, k), \, |k| \le 2^{j+1} |\sin\theta_{j,\ell}|, \text{ such that (s.t.) } |b_{j,\ell,k}| \ge \epsilon\},$$

we want to show that

$$\sup_{\epsilon > 0} \epsilon N^{1/p^*}(\epsilon) \le C \, \|g\|_{H^s}$$

since the left-hand side is an equivalent definition of the weak-ℓ_{p^*} norm (1.5).

Put

$$N_{j,m}(\epsilon) = \#\{(\ell,k),\, \ell \in \Lambda_{j,m},\, |k| \le 2^{j+1}|\sin\theta_{j,\ell}|, \text{ s.t.} |b_{j,\ell,k}| \ge \epsilon\}.$$

Corollary 4.2 posits the existence of a constant K such that $|b_{j,\ell,k}|^2 \le K\, 2^{-j}\|g\|^2_{H^s}$ (4.3) and therefore, it is clear that $N_{j,m}(\epsilon) = 0$ if $2^j \ge K\,\epsilon^{-2}\|g\|^2_{H^s}$. In what follows, we will let η be defined by $\eta = \epsilon/\|g\|_{H^s}$. Regardless of the condition $|b_{j,\ell,k}| \ge \epsilon$, the cardinality of the index set $\{(\ell,k),\, \ell \in \Lambda_{j,m},\, |k| \le 2^{j+1}|\sin\theta_{j,\ell}|\}$ is bounded by $C\,2^{d(j-m)}$. Further, the bound on the ℓ_2 norm of the $b_{j,\ell,k}$'s (Corollary 4.2) gives

$$N_{j,m}(\epsilon) \le C\, \min(2^{(j-m)d}, \eta^{-2}2^{-j}2^{(j-m)(1-2s)})$$

whenever $2^j \le K\,\eta^{-2}$.

Let $N_j(\epsilon)$ be the number of coefficients whose absolute values exceed ϵ, i.e.,

$$N_j(\epsilon) = \#\{(\ell,k),\, |k| \le 2^{j+1}|\sin\theta_{j,\ell}|,\, |b_{j,\ell,k}| \ge \epsilon\}.$$

Then, a simple calculation gives

$$\begin{aligned} N_j(\epsilon) = \sum_m N_{j,m}(\epsilon) &\le C \sum_m \min(2^{(j-m)d}, \eta^{-2}2^{-j}2^{(j-m)(1-2s)}) \\ &\le C\, \min(2^{jd}, \eta^{-2d/\alpha}2^{-jd/\alpha}), \end{aligned}$$

where $\alpha = d + 2s - 1$. To summarize, we have

$$N_j(\epsilon) \le C \begin{cases} 0 & 2^j \ge K\,\eta^{-2}, \\ \eta^{-2d/\alpha}2^{-jd/\alpha} & \eta^{-2/(1+\alpha)} \le 2^j \le K\,\eta^{-2}, \\ 2^{jd} & 2^j \le \eta^{-2/(1+\alpha)}. \end{cases}$$

Summing over the scales yields

$$\begin{aligned} N(\epsilon) = \sum_{j=0}^{\infty} N_j(\epsilon) &\le C \sum_{j:2^j \le \eta^{-2/(1+\alpha)}} 2^{jd} + C \sum_{j:\eta^{-2/(1+\alpha)} \le 2^j \le K\,\eta^{-2}} \eta^{-2d/\alpha}2^{-jd/\alpha} \\ &\le C\,\eta^{-2d/(1+\alpha)} = C\,\eta^{-p^*} = C\,\epsilon^{-p^*}\,\|g\|^{p^*}_{H^s} \end{aligned}$$

with $1/p^* = s/d + 1/2$. This finishes the proof of the first assertion.

We now turn to the second assertion. It clearly follows from (4.4) that for any $q > 0$ we have

$$\sum_{k:|k|>2^{j+1}|\sin\theta_{j,\ell}|} |b_{j,\ell,k}|^q \le C\,2^{jq/2}(2^j|\sin\theta_{j,\ell}|)^{1-nq}\|g\|^q_{H^s},$$

since n may be chosen arbitrarily large and, in particular, greater than $1/q$. Summing over the ℓ's, $\ell \in \Lambda_{j,m}$ gives

$$\sum_{\ell\in\Lambda_{j,m}} \sum_{k:|k|>2^{j+1}|\sin\theta_{j,\ell}|} |b_{j,\ell,k}|^q \le C\,2^{jq/2}2^{(1-nq)(j-m)}2^{(j-m)(d-1)}\|g\|^q_{H^s}.$$

Now, we must keep in mind that we have available a bound on the ℓ_2 norm (4.3); i.e.,

$$\sum_{\ell\in\Lambda_{j,m}} \sum_{k:|k|>2^{j+1}|\sin\theta_{j,\ell}|} |b_{j,\ell,k}|^2 \le C\,2^{-j}2^{-(j-m)(2s-1)}\|g\|^2_{H^s}.$$

The interpolation inequality will yield the ℓ_p boundedness. Recall that for any sequence a_n we have

$$\|a\|_{\ell_p} \le \|a\|_{\ell_q}^{\theta} \, \|a\|_{\ell_2}^{1-\theta}, \quad 1/p = \theta/q + (1-\theta)/2. \tag{4.12}$$

This interpolation inequality applied to our subsequence gives

$$\left(\sum_{\ell \in \Lambda_{j,m}} \sum_{k:|k|>2^{j+1}|\sin\theta_{j,\ell}|} |b_{j,\ell,k}|^p \right)^{1/p} \le C \left[2^{j/2} 2^{-(j-m)(n-d/q)} \right]^{\theta} \left[2^{-j/2} 2^{-(j-m)(s-1/2)} \right]^{1-\theta} \|g\|_{H^s}.$$

In the previous inequality, the value of n may be chosen arbitrarily large and, hence, summing up the previous inequalities results in the upper bound

$$\sum_{\ell} \sum_{k:|k|>2^{j+1}|\sin\theta_{j,\ell}|} |b_{j,\ell,k}|^p \le C \, 2^{-jp(1/2-\theta)} \, \|g\|_{H^s}^p. \tag{4.13}$$

This establishes the boundedness in ℓ_p for any $p > 0$. Indeed for $p > 0$, choose q small enough so that $\theta < 1/2$ (4.12), i.e., $1/q > 2/p + 1/2$, and apply (4.13). The theorem is proved for $s = 1, 2, \ldots$.

Interpolation theory allows us to extend the result to the half line $s > 0$. Indeed, let T be the operator

$$T : \, g \mapsto (\alpha_\nu)$$

that maps g into the ridgelet coefficient sequence (α_ν) of f, $f(x) = H(u \cdot x - b)g(x)$, with u and b fixed. We abuse notations—as it is understood that we are concerned with elements supported on the unit cube—and let H^s be the Banach space defined by

$$H^s := \{g, \; g \in H^s \text{ and supp } g \subset [0,1]^d\}$$

equipped with the norm $\| \cdot \|_{H^s}$. We proved that for any $n \ge 1$, $\|T\|$ is a bounded operator from H^n to $w\ell_p$, $1/p = n/d + 1/2$. In addition, T is bounded from L_2 to ℓ_2 (where again we understand $L_2([0,1]^d)$). On one hand, it is well known that (L_2, H^n) is an interpolation couple [2] and that for any $n > 0$ and any $0 < \theta < 1$, we have

$$(L_2, H^n)_{\theta,2} = H^{n\theta};$$

see [14], for example. On the other, letting ℓ_2 be the space of real-valued sequences

$$\ell_2 = \left\{ a, \; \sum_{n \ge 1} |a_n|^2 < \infty \right\},$$

and similarly for $w\ell_p$, $p > 0$, we have

$$(\ell_2, w\ell_p)_{\theta,2} = \ell_{p^*,2}, \quad 1/p^* = (1-\theta)/2 + \theta/p.$$

Here, $\ell_{p,2}$, $p > 0$ is the Lorentz space of real sequences

$$\left(\sum_{n \ge 1} |a|_{(n)}^2 n^{2/p-1} \right)^{1/2} < \infty,$$

where we recall that $|a|_{(n)}$ is the nth largest entry in the sequence $(|a_n|)$. The interpolation theorem [2] gives that

$$T : H^{n\theta} \to \ell_{p^*,2}$$

is bounded and further that

$$\|T\|_{H^{n\theta}\to\ell_{p^*,2}} \leq C\, \|T\|^{1-\theta}_{L_2\to\ell_2} \|T\|^{\theta}_{H^n\to w\ell_p}.$$

Hence, for any $s > 0$, pick $n > s$ and put $\theta = s/n$. We have

$$\frac{1}{p^*} = \frac{1}{2}\left(1 - \frac{s}{n}\right) + \frac{s}{n}\left(\frac{n}{d} + \frac{1}{2}\right) = \frac{s}{d} + \frac{1}{2},$$

and, therefore, our analysis gives that T is bounded from H^s to $\ell_{p^*,2}$. This completes the proof of our theorem since for any sequence a and any $p > 0$, we have

$$\|a\|_{\ell_{p,2}} \leq \|w\ell_p\|. \qquad \square$$

Remark. We proved a slightly stronger result than that announced in our theorem since for any $s \geq 0$ the ridgelet coefficient sequence obeys

$$\|\alpha\|_{\ell_{p,2}} \leq C\, \|g\|_{H^s},\ 1/p = s/d + 1/2.$$

4.1. Finite approximations. We now exploit Theorem 4.1 to derive nonlinear approximation bounds. The compact notation $(\psi_\nu)_{\nu\in\mathcal{N}}$ introduced in section 2 will be used to denote the frame elements.

Suppose that f is of the form

$$f(x) = g_0(x) + H(u \cdot x - b) g_1(x), \tag{4.14}$$

where

$$\|g_i\|_{H^s} \leq C, \quad i = 0, 1.$$

From the exact series

$$f = \sum_{\nu\in\mathcal{N}} \alpha_\nu \widetilde{\psi}_\nu,$$

extract the n-term approximation f_n obtained by keeping the n terms corresponding to the n largest coefficients. Then, we have the following result.

Corollary 4.3. *With the previous assumptions, there exists a constant C (not depending on f) such that*

$$\|f - f_n\|_2 \leq C\, n^{-s/d} \sup_{i=0,1} \|g_i\|_{H^s(\mathbb{R}^d)}. \tag{4.15}$$

As we will see below, the convergence rate of n-term ridgelet approximations is, in some sense, optimal.

Theorem 4.1 gives that the coefficients (α_ν) of f are bounded in $w\ell_{p^*}$. Letting $|\alpha|_{(n)}$ be the nth largest entry in α (in absolute values), we have

$$f - f_n = \sum_\nu \alpha_\nu 1_{\{|\alpha_\nu| \geq |\alpha|_{(n)}\}} \widetilde{\psi}_\nu.$$

The lemma stated below then gives the desired conclusion, namely,

$$\|f - f_n\|_2^2 \le A^{-1} \sum_{m>n} |\alpha|_{(m)}^2 \le A^{-1}\, C\, n^{-2s/d} \|\alpha\|_{w\ell_{p^*}}^2,$$

where A is the constant appearing on the left-hand side of (2.7).

LEMMA 4.4. *Let $(a_\nu)_{\nu\in\mathcal{N}}$ be a sequence in ℓ_2 and let*

$$\tilde{f} = \sum_{\nu\in\mathcal{N}} a_\nu \tilde{\psi}_\nu.$$

Then,

$$\|\tilde{f}\|_2^2 \le A^{-1} \|a\|_{\ell_2}^2.$$

Proof. We let $\tilde{F}$ be the synthesis operator defined by $\tilde{F}a = \sum a_\nu \tilde{\psi_\nu}$ and let F be the analysis operator $Ff = (\langle f, \psi_\nu\rangle)_{\nu\in\mathcal{N}}$. The property (2.7) gives

$$\|\tilde{f}\|^2 = \|\tilde{F}a\|^2 \le A^{-1} \|F\,\tilde{F}a\|_{\ell_2}^2.$$

Now, it is easy to see that $F\,\tilde{F}$ is the orthogonal projector onto the range of F and has, therefore, a norm (as an operator from ℓ_2 onto itself) bounded by 1. Consequently, we have

$$\|\tilde{f}\|^2 \le A^{-1} \|F\,\tilde{F}a\|_{\ell_2}^2 \le A^{-1} \|a\|_{\ell_2}^2,$$

which is what needed to be shown. □

4.2. Optimality. In this section, we detail the sense in which Corollary 4.3 is optimal. Consider a class of templates of the form (4.14): i.e., let $\mathcal{F}(C)$ be the class defined by

$$\mathcal{F}(C) = \{f,\ f \text{ satisfies (4.14)},\ \|g_i\|_{H^s} \le C, \text{ and } \operatorname{supp} g_i \subset [0,1]^d,\ i = 0,1\}. \tag{4.16}$$

In the above definition, the singular hyperplane is not fixed; two elements from $\mathcal{F}(C)$ may be singular along two different hyperplanes.

The class $\mathcal{F}(C)$ contains, of course, the Sobolev ball $H^s(C) = \{f,\ \|f\|_{H^s} \le C, \text{ and } \operatorname{supp} f \subset [0,1]^d\}$. In any orthobasis $(\phi)_{i\in\mathcal{I}}$, there is a lower bound on the convergence of the best n-term approximation $Q_n(f)$ in that basis,

$$\sup_{f\in H^s(C)} \|f - Q_n(f)\|_2 \ge C\, n^{-s/2}.$$

As a consequence, no orthobasis exists that provides better rates than those obtained in Corollary 4.3. There is even a broader notion of optimality based on information theoretic concepts such as the Kolmogorov ϵ-entropy or the minimum description length (MDL) paradigm.

Let $\mathcal{F}$ be a compact set of functions in $L^2([0,1]^d)$. The Kolmogorov ϵ-entropy $N(\epsilon, \mathcal{F})$ of the class $\mathcal{F}$ is the minimum number of bits that is required to specify any element f from $\mathcal{F}$ within an accuracy of ϵ. In other words, let ℓ be a fixed counting number and let $E_\ell :\ \mathcal{F} \to \{0,1\}^\ell$ be a functional which assigns a bit string of length ℓ to each $f \in \mathcal{F}$. Let $D_\ell :\ \{0,1\}^\ell \to L_2[0,1]^d$ be a mapping which assigns to each bit string of length ℓ a function. The coder-decoder pair (E_ℓ, D_ℓ) will be said to achieve a distortion $\leq \epsilon$ over $\mathcal{F}$ if

$$\sup_{f\in\mathcal{F}} \|D_\ell(E_\ell(f)) - f\| \leq \epsilon.$$

The Kolmogorov ϵ-entropy (minimax description length) may then be defined as

$$L^*(\epsilon, \mathcal{F}) = \min\{\ell :\ \exists (E_\ell, D_\ell) \text{ achieving distortion } \leq \epsilon \text{ over } \mathcal{F}\}.$$

The minimum number of bits needed to reconstruct any f taken from our class of templates $\mathcal{F}(C)$ (4.16) satisfies

$$N(\epsilon, \mathcal{F}(C)) \geq N(\epsilon, H^s) \geq C\, \epsilon^{2/s}.$$

A strategy identical to that developed in [9, Theorem 2], however, gives a simple way to exploit the sparsity of the ridgelet sequence to construct a coder-decoder pair of length $O(\log(\epsilon^{-1})\epsilon^{2/s})$ that achieves a distortion of ϵ. The construction is based on simple uniform quantization of the ridgelet coefficients α_i, followed by simple run length coding. Hence, we have available a very concrete way of obtaining near-optimal (possibly within log-like factors) compression rates.

5. Orthonormal ridgelets. In dimension 2, Donoho [10] introduced a new orthonormal basis whose elements he called "orthonormal ridgelets." We will not detail why these elements relate to ridgelets. We quote from [7]: "Such a system can be defined as follows: let $(\psi_{j,k}(t) : j \in \mathbb{Z}, k \in \mathbb{Z})$ be an orthonormal basis of Meyer wavelets for $L^2(\mathbb{R})$ [12], and let $(w^0_{i_0,\ell}(\theta),\ \ell = 0, \dots, 2^{i_0}-1;\ w^1_{i,\ell}(\theta),\ i \geq i_0,\ \ell = 0, \dots, 2^i-1)$ 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.) 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

$$\hat{\rho}_\lambda(\xi) = |\xi|^{-\frac{1}{2}}(\hat{\psi}_{j,k}(|\xi|)w^\varepsilon_{i,\ell}(\theta) + \hat{\psi}_{j,k}(-|\xi|)w^\varepsilon_{i,\ell}(\theta + \pi))/2\ . \tag{5.1}$$

Here the indices run as follows: $j, k \in \mathbb{Z}$, $\ell = 0, \dots, 2^{i-1} - 1$; $i \geq i_0$, $i \geq j$. Notice the restrictions on the range of ℓ and on i. Let λ denote the set of all such indices λ. It turns out that $(\rho_\lambda)_{\lambda\in\Lambda}$ is a complete orthonormal system for $L^2(\mathbb{R}^2)$."

There is a close connection between "pure" and orthonormal ridgelets. Pure ridgelets are supported on lines in the Fourier domain: that is, the frequency representation of a pure ridgelet is given by (provided that the profile ψ is real-valued)

$$\hat{\psi}_{j,\ell,k}(\xi) = (\hat{\psi}_{j,k}(|\xi|)\delta(\theta - 2\pi 2^{-j}\ell) + \hat{\psi}_{j,k}(-|\xi|)\delta(\theta + \pi - 2\pi 2^{-j}\ell))/2 \tag{5.2}$$

using a formulation emphasizing the resemblance with (5.1). In the ridgelet construction, the angular variable θ is uniformly sampled at each scale, the sampling step being inversely proportional to the scale. In contrast, the sampling idea is replaced by the wavelet transform for orthonormal ridgelets. This is the reason why orthonormal ridgelets can perfectly reconstruct objects from $L^2(\mathbb{R}^2)$ without support constraints. It is interesting to note that the restriction on the range, namely, $i \geq j$ in the definition (5.1), gives angular scaling functions at scales inversely proportional to the sampling steps of pure ridgelets.

THEOREM 5.1. *Let $g \in H^s(\mathbb{R}^2)$, $s > 0$, with compact support and put $f(x) = H(u \cdot x - b)\, g(x)$. Then the orthonormal ridgelet coefficient sequence α of f obeys*

$$\|\alpha\|_{w\ell_p} \le C\, \|g\|_{H^s} \quad \text{with} \quad 1/p = s/2 + 1/2$$

for some constant C not depending on f. It then follows that the truncated n-term partial reconstruction f_n achieves the error bound

$$\|f - f_n\|_2 \le C\, n^{-s/2} \|g\|_{H^s}.$$

Proof. The proof is an application of Theorem 3.1 and consists of minor modifications to the proof of Theorem 4.1. In the following, we outline the essential steps, thus avoiding worthless repetition.

Begin with $\varepsilon = 0$ $(i = j)$ and observe that

$$|\langle f, \rho_\lambda \rangle| = \left| \int \hat{f}(\lambda, \theta)\, |\lambda|^{1/2} (\hat{\psi}_{j,k}(|\lambda|) w_{j,\ell}^{\varepsilon=0}(\theta) + \hat{\psi}_{j,k}(-|\lambda|) w_{j,\ell}^{\varepsilon=0}(\theta + \pi))\, d\lambda d\theta \right| /2$$

$$\le 2^{j/2} \int |w_{j,\ell}^{\varepsilon=0}(\theta)| J^+(\theta) d\theta/2 + 2^{j/2} \int |w_{j,\ell}^{\varepsilon=0}(\theta + \pi)| J^-(\theta) d\theta/2, \tag{5.3}$$

where

$$J^{\pm}(\theta) = \left| \int \hat{f}(\lambda, \theta)\, |2^{-j}\lambda|^{1/2} \hat{\psi}_{j,k}(\pm|\lambda|) d\lambda \right|.$$

The point of this paper has been precisely to bound quantities like $J^{\pm}(\theta)$. For instance, let $I_{j,\ell} = \{\theta, |\theta - 2\pi\, 2^{-j}\ell| \le 2^{-j}\}$ and set

$$\beta_{j,\ell,k} = 2^j \int_{I_{j,\ell}} \left| \int \hat{f}(\lambda, \theta) |2^{-j}\lambda|^{1/2} \hat{\psi}_{j,k}(|\lambda|) d\lambda \right|.$$

Then, we proved that (dimension 2)

$$\|\beta\|_{w\ell_p} \le C\, \|g\|_{H^s}, \quad 1/p = s/2 + 1/2.$$

Compare the previous inequality with (4.5) and Theorem 4.1. Second, the scaling function is localized near the interval $I_{j,\ell}$; for any $\gamma > 0$, there is a constant C such that

$$|w_{j,\ell}^{\varepsilon=0}(\theta)| \le C\, 2^{j/2} (1 + 2^j |\theta - 2\pi\, \ell 2^{-j}|)^{-\gamma}.$$

Hence, a reasoning similar to the one developed for Theorem 4.1 gives

$$\|\alpha^{\varepsilon=0}\|_{w\ell_p} \le C\, \|g\|_{H^s}, \quad 1/p = s/2 + 1/2. \tag{5.4}$$

The point is that the contributions associated with the orthonormal ridgelets corresponding to parameter values $i > j$ become negligible as i goes to infinity. This is due to the compactness of the support of f. Letting D be $\partial/\partial\theta$, standard wavelet calculations give

$$\langle f, \rho_\lambda \rangle = I_n^+ + I_n^-,$$

$$I_n^+ = \int D^n \hat{f}(\lambda, \theta)\, |\lambda|^{1/2} (\hat{\psi}_{j,k}(|\lambda|) D^{-n} w_{i,\ell}^{\varepsilon=1}(\theta)\, d\theta,$$

and similarly for I_n^-. Both terms are treated identically. Since

$$|D^{-n} w_{i,\ell}^{\varepsilon=1}(\theta)| \le C\, 2^{-i(n-1/2)} (1 + 2^i|\theta - 2\pi\,\ell 2^{-i}|)^{-\gamma}$$

we have

$$|I_n^+| \le C\, 2^{-in} 2^{i/2} 2^{j/2} \int (1 + 2^i|\theta - 2\pi\,\ell 2^{-i}|)^{-\gamma} J_n^+(\theta)\, d\theta,$$

where now

$$J_n^{\pm}(\theta) = \left| \int (\partial_\theta^n \hat{f})(\lambda, \theta)\, |2^{-j}\lambda|^{1/2} \hat{\psi}_{j,k}(\pm|\lambda|) d\lambda \right|.$$

Observe now that

$$\partial_\theta \hat{f}(\lambda, \theta) = \lambda(-\sin\theta(\partial_1 \hat{f})(\lambda, \theta) + \cos\theta(\partial_2 \hat{f})(\lambda, \theta)),$$

and this formula may be iterated to obtain derivatives with respect to the angular variable θ of higher orders.

We may then substitute polar derivatives with respect to θ by cartesian derivatives and obtain (letting D be either $\partial/\partial x_1$ or $\partial/\partial x_2$)

$$|I_n^+| \le C\, 2^j 2^{-(i-j)(n-1/2)} \int (1 + 2^i|\theta - 2\pi\,\ell 2^{-i}|)^{-\gamma} \sum_{|\alpha| \le n} J_\alpha^+(\theta)\, d\theta,$$

$$J_\alpha^+(\theta) = \left| \int (D^\alpha \hat{f})(\lambda, \theta)\, |2^{-j}\lambda|^{|\alpha|+1/2} \hat{\psi}_{j,k}(|\lambda|) d\lambda \right|.$$

We already argued in the proof of Corollary 4.2 that, because of the compactness of the support of the distribution f, the estimates we obtained for $\hat{f}$ are valid for the derivatives $D^\alpha \hat{f}$. Hence, we essentially have the same bound as in (5.3) but for an exponentially decaying factor $2^{-(i-j)(n-1/2)}$, where n might be chosen as large as we want. It is then not too difficult to check that the sequence $\alpha^{\varepsilon=1}$ satisfies

$$\|\alpha^{\varepsilon=1}\|_{w\ell_p} \le C\, \|g\|_{H^s}, \quad 1/p = s/2 + 1/2.$$

The $w\ell_p$ boundedness of the sequence α naturally follows from this last display and (5.4). □

6. Discussion. Unlike any known system, ridgelets allow optimal partial reconstructions of L_2 Sobolev functions with linear singularities. These good approximations are, moreover, simply obtained by thresholding the exact ridgelet series (1.4).

6.1. Ridgelets and functional classes. As we pointed out in the introduction, wavelets are optimal to represent smooth functions with point-singularities. From a functional viewpoint, we may say that wavelets provide unconditional bases for the Besov spaces and the Triebel spaces [13] and, therefore, they provide near-optimal approximations to elements taken from functional balls of such spaces. A natural question would be, What are the functional spaces that are naturally associated with ridgelets? The analysis that we presented already suggests an answer. It is certainly possible to build new functional spaces whose typical elements resemble our mutilated

Sobolev objects. In this direction, we might be tempted to consider, for instance, convex combinations of objects like (1.2); let

$$\mathcal{S}_H = \left\{ f = \sum_i a_i f_i, \ \sum_i |a_i| \le 1 \right\},$$

where the f_i's are our templates, i.e., functions of the form

$$f_i(x) = H(u_i \cdot x - b_i) g_i(x), \quad \|g_i\|_{H^s} \le 1, \text{ supp } g \subset [0,1]^d.$$

Our functional class $\mathcal{S}_H$ would then be meant to represent objects composed of singularities across hyperplanes: typical elements of this class are discontinuous across these same hyperplanes and otherwise smooth. There may be an arbitrary number of singularities which may be located in all orientations and positions. In the author's unpublished thesis [3], it is then proved that ridgelets provide near-optimal representations of objects of this kind, as expected.

This is, indeed, part of a larger picture. A new notion of smoothness may be introduced leading to new functional classes that are naturally associated with ridgelets. This new notion of smoothness is nonclassical; it is discussed in [3] and briefly exposed in [7]. Full details will be provided in a separate paper.

6.2. Curved singularities. We would like to emphasize that this paper considered only linear singularities. Ridgelets are not able to efficiently represent smooth functions with curved singularities. For instance, in dimension d, consider the indicator function of the unit ball

$$f(x) = 1_{\{|x| \le 1\}},$$

and let α denote the ridgelet coefficient sequence of f. Then, [3] shows that

$$\#\{n, \text{ s.t.} \, |\alpha_n| \ge 1/n\} \ge C \, n^{2(1-1/d)}, \tag{6.1}$$

yielding partial reconstructions converging only at the rate $n^{-\frac{1}{2(d-1)}}$. We quote from [7]: "Unfortunately, the task that ridgelets must face is somewhat more difficult than the task which wavelets must face, since zero-dimensional singularities are inherently simpler objects than higher-dimensional singularities. In effect, zero-dimensional singularities are all the same—points—while a one-dimensional singularity—lying along a one-dimensional set—can be curved or straight." It is remarkable, however, that both wavelet and ridgelets, two fundamentally different systems, achieve the same degree of sparsity.

The method of localization enables us to obtain sharper approximation bounds on objects with curved singularities. The localization idea is rather straightforward and has been, for instance, previously deployed in the time frequency literature. We outline this idea in dimension 2: first, partition the unit square into small squares, and smoothly localize the function into smooth pieces supported on or near those squares; then take the ridgelet transform on each piece. This is the basis of the so-called *monoscale ridgelet transform* [5]. Again, partial reconstructions simply obtained by keeping the largest coefficients are shown to provide good approximation bounds (of higher order than wavelet or ridgelet approximations).

Further, [6] developed a new approach, namely, the *curvelet transform* that combines ideas from ridgelet analysis and wavelet analysis. In two dimensions, the curvelet transform provides optimal representations of smooth functions with twice differentiable singularities, a fact whose roots are grounded on the results presented in this paper.

Acknowledgments. I am especially grateful to David Donoho for many fruitful discussions. I would also like to thank one referee for some very helpful comments on the original version of the manuscript.

Some of the results were briefly described at the Royal Society meeting "Wavelets: a key to intermittent information?" held in London in February 1999.

REFERENCES

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

[2] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction*, Grundlehren Math. Wiss. 223, Springer-Verlag, Berlin, New York, 1976.

[3] E. J. Candes, *Ridgelets: Theory and Applications*, Ph.D. thesis, Department of Statistics, Stanford University, Stanford, CA, 1998.

[4] E. J. Candes, *Harmonic analysis of neural netwoks*, Appl. Comput. Harmon. Anal., 6 (1999), pp. 197–218.

[5] E. J. Candes, *Monoscale Ridgelets for the Representation of Images with Edges*, Tech. report, Department of Statistics, Stanford University, Stanford, CA, 1999.

[6] E. J. Candès and D. L. Donoho, *Curvelets—a surprisingly effective nonadaptive representation for objects with edges*, in Curves and Surfaces, A. Cohen, C. Rabut, and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 1999.

[7] E. J. Candes and D. L. Donoho, *Ridgelets: The key to higher-dimensional intermittency?*, Philos. Trans. Roy. Soc. London Ser. A., 357 (1999), pp. 2495–2509.

[8] D. L. Donoho, *Unconditional bases are optimal bases for data compression and for statistical estimation*, Appl. Comput. Harmon. Anal., 1 (1993), pp. 100–115.

[9] D. L. Donoho, *Unconditional bases and bit-level compression*, Appl. Comput. Harmon. Anal., 3 (1996), pp. 388–392.

[10] D. L. Donoho, *Orthonormal ridgelets and linear singularities*, SIAM J. Math. Anal., 31 (2000), pp. 1062–1099.

[11] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies, *Data compression and harmonic analysis*, IEEE Trans. Inform. Theory, 44 (1998), pp. 2435–2476.

[12] P. G. Lemarié and Y. Meyer, *Ondelettes et bases Hilbertiennes*, Rev. Mat. Iberoamericana, 2 (1986), pp. 1–18.

[13] Y. Meyer, *Wavelets and Operators*, Cambridge University Press, Cambridge, UK, 1992.

[14] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1978.

[15] H. Triebel, *Theory of Function Spaces.* II, Monogr. Math. 84, Birkhäuser-Verlag, Basel, 1992.

SIAM J. Math. Anal.
Vol. 33, No. 2, pp. 369–389

AN APPROXIMATION SCHEME FOR MOTION BY MEAN CURVATURE WITH RIGHT-ANGLE BOUNDARY CONDITION*

HITOSHI ISHII† AND KATSUYUKI ISHII‡

Dedicated to the late Professor Yoshihito Tomita

Abstract. We show that the algorithm considered by Ishii [*GAKUTO Internat. Ser. Math. Sci. Appl.* 5, Gakkōtosho, Tokyo, 1995, pp. 111–127] and Ishii, Pires, and Souganidis [*J. Math. Soc. Japan,* 50 (1999), pp. 267–308] can be applied to motion by mean curvature with right-angle boundary condition.

Key words. motion by mean curvature, right-angle boundary condition, approximation scheme, viscosity solutions

AMS subject classifications. 35K65, 35K55, 65M99

PII. S0036141000368107

1. Introduction.

In 1992, Bence, Merryman, and Osher proposed an algorithm for computing motion of a hypersurface by mean curvature (cf. [2]). It is described as follows. For a given closed set $C_0 \subset \mathcal{R}^N$, let u be a solution of the initial value problem for the heat equation

$$\left\{ \begin{array}{lll} u_t - \Delta u = 0 & \text{in} & (0, +\infty) \times \mathcal{R}^N, \\ u(0,x) = \chi_{C_0}(x) & \text{in} & \mathcal{R}^N, \end{array} \right. \tag{1.1}$$

where χ_{C_0} is the characteristic function of C_0. Fix a time step $h > 0$ and set

$$C_1 = \left\{ x \in \mathcal{R}^N \;\middle|\; u(h,x) \geqq \frac{1}{2} \right\}. \tag{1.2}$$

Next we solve (1.1) with C_0 replacing C_1 and define C_2 as the set in (1.2) with u replaced by the solution with the new initial data. Repeating this procedure, we obtain a sequence $\{C_k\}_{k \in \mathcal{N}}$ of closed subsets in $\mathcal{R}^N$. Then we set

$$C_t^h = C_k \qquad \text{if } kh \leqq t < (k+1)h \text{ and } k \in \mathcal{N} \cup \{0\}$$

for $t \geqq 0$. Letting $h \to 0$, we obtain in the limit a flow $\{M_t\}_{t \geqq 0}$ of closed subsets in $\mathcal{R}^N$ with $M_0 = C_0$, whose boundary moves by ($(n-1)$-times) mean curvature.

The convergence of this algorithm was proved by Evans [4] and Barles and Georgelin [1]. Noticing that the solution of (1.1) is given by taking the convolution with Gauss kernel and χ_{C_0}, Ishii [6] extended this algorithm to the case of general radially symmetric kernels. Ishii, Pires, and Souganidis [7] considered threshold dynamics type approximation schemes, which are also an extension of the above algorithm.

*Received by the editors February 22, 2000; accepted for publication (in revised form) February 8, 2001; published electronically July 19, 2001.
http://www.siam.org/journals/sima/33-2/36810.html

†Department of Mathematics, School of Education, Waseda University, Nishi-Waseda 1-6-1, Shijuku, Tokyo 169-8050, Japan (hishii@comp.metro-u.ac.jp). This author was partially supported by Grant-in-Aid for Scientific Research 09440067, from the Ministry of Education, Science, Sports, and Culture.

‡Department of Mathematics, Kobe University of Mercantile Marine, Higashinada, Kobe 658-0022, Japan (ishii@cc.kshosen.ac.jp). This author was partially supported by Grant-in-Aid for Scientific Research 10640119, from the Ministry of Education, Science, Sports, and Culture.

The main purpose of this paper is to show that the scheme introduced by [6] and [7] can be applied to the motion by mean curvature with right-angle boundary condition in a bounded domain. Our analysis is based on the level set approach as in [4], [6], and [7].

This paper is organized in the following way. In section 2 we recall briefly the level set approach to motion by mean curvature with right-angle boundary condition and introduce our scheme. Also, we state our main result in section 2. Section 3 is devoted to the estimates for the operator G_h. The key idea is to compare G_h with $\widetilde{G}_h$ defined in $\mathcal{R}^N$. However, our arguments are complicated because we need some estimates of the integration of $f_{\sqrt{h}}(\cdot - x)$ outside $B_N(x, \sqrt{h}R(\sqrt{h}))$. Thus we divide section 3 into two subsections. In section 3.1, we consider the case where the support of the kernel function f is compact. Thanks to the compactness of supp$f$, the arguments become easy and understandable. In section 3.2 we treat the noncompact case. We prove our main result in section 4.

2. Preliminaries and the main result.

First let us recall briefly the level set approach to motion by mean curvature with right-angle boundary condition. See Sato [8] and Giga and Sato [5] for the details.

Let $\Omega \subset \mathcal{R}^N$ be a bounded domain with C^2 boundary $\partial\Omega$ and let $g \in C(\overline{\Omega})$. We consider the initial-boundary value problem:

$$(2.1) \qquad \begin{cases} u_t + F(Du, D^2u) = 0 & \text{in} \quad (0, +\infty) \times \Omega, \\ \dfrac{\partial u}{\partial n} = 0 & \text{in} \quad (0, +\infty) \times \partial\Omega, \\ u(0, x) = g(x) & x \in \overline{\Omega}, \end{cases}$$

where $F(p, X) = -\mathrm{tr}\{(I - p \otimes p/|p|^2)X\}$ and n denotes the outer unit normal to $\partial\Omega$. Problem (2.1) is the level set equation of the mean curvature flow with right-angle boundary condition since each level set of (2.1) moves by its mean curvature in Ω and it intersects $\partial\Omega$ perpendicularly, at least formally. The above equation is degenerate parabolic and has singularities for $Du = 0$. In spite of the difficulties coming from these facts, problem (2.1) has a unique viscosity solution in $C([0, T) \times \overline{\Omega})$ for any $T > 0$. Moreover, it has been shown that, for the viscosity solution u of (2.1), the level set $\{x \in \overline{\Omega} \mid u(t, x) = c\}$ is determined by the set $\{x \in \overline{\Omega} \mid g(x) = c\}$ and is independent of the choice of g. Also, it has been shown that if $\{M_t\}_{0 \leqq t < T}$ is a flow of a closed subset of $\overline{\Omega}$ such that $\{\partial M_t\}_{0 \leqq t < T}$ is a smooth flow by mean curvature satisfying $\partial M_t \perp \partial\Omega$ for all $t \in [0, T)$ and if u is a viscosity solution of (2.1) with $\{x \in \overline{\Omega} \mid g(x) = c\} = \partial M_0$ (resp., $\{x \in \overline{\Omega} \mid g(x) \geqq c\} = M_0$), then $\{x \in \overline{\Omega} \mid u(t, x) = c\} = \partial M_t$ (resp., $\{x \in \overline{\Omega} \mid u(t, x) \geqq c\} = M_t$). Thus the family $\{x \in \overline{\Omega} \mid u(t, x) = c\}_{0 \leqq t < T}$ is often called a generalized motion by mean curvature with right-angle boundary condition.

Now we formulate an approximation scheme for motion by mean curvature with right-angle boundary condition, according to Ishii [6] and Ishii, Pires, and Souganidis [7]. Let f be a real-valued function on $\mathcal{R}^N$. We make the following assumptions.

(A.1) f is nonnegative, measurable, and radially symmetric.

(A.2) $\int_{\mathcal{R}^N} f(x)(1 + |x|^2)dx < +\infty$.

(A.3) $\int_{\mathcal{R}^{N-1}} f(\xi, 0)(1 + |\xi|^2)d\xi < +\infty$.

(A.4) Let $\{R(\rho)\}_{0<\rho<1}$ satisfy

$$R(\rho) \to +\infty, \rho R(\rho)^2 \to 0 \quad (\rho \to 0).$$

Then for any $g(\xi) = \langle A\xi, \xi\rangle + a$ ($A \in \mathcal{S}^{N-1}$, $a \in \mathcal{R}$),

$$\lim_{\rho\to 0} \sup_{0<r<\rho} \left| \int_{B_{N-1}(0,R(\rho))} f(\xi, rg(\xi))g(\xi)d\xi - \int_{\mathcal{R}^{N-1}} f(\xi, 0)g(\xi)d\xi \right| = 0.$$

We introduce the operator $G_h : C(\overline{\Omega}) \to C(\overline{\Omega})$ with $h > 0$ by

$$G_h g(x) = \sup\left\{ \lambda \in \mathcal{R} \;\middle|\; \int_\Omega f_{\sqrt{h}}(y-x)\chi_{\{g \geqq \lambda\}}(y)dy \geqq \frac{1}{2}\int_\Omega f_{\sqrt{h}}(y-x)dy \right\}$$

for $g \in C(\overline{\Omega})$. Here and in the sequel $f_{\sqrt{h}}(x) = h^{-N/2} f(x/\sqrt{h})$, $\{g \geqq \lambda\} = \{y \in \overline{\Omega} \mid g(y) \geqq \lambda\}$, and χ_A is the characteristic function of $A \subset \mathcal{R}^N$. It is easily seen that, for any g_1, g_2, $g \in C(\overline{\Omega})$,

$$G_h g_1 \leqq G_h g_2 \quad \text{if } g_1 \leqq g_2, \tag{2.2}$$

$$G_h(g+c) = G_h g + c, G_h c = c \quad \text{for all } c \in \mathcal{R}, \tag{2.3}$$

$$G_h(\theta \circ g) = \theta \circ G_h g \quad \text{for any nondecreasing function } \theta \in C(\mathcal{R}). \tag{2.4}$$

It follows from (2.2)–(2.4) that the operator G_h is contractive and nonexpansive on $C(\overline{\Omega})$.

Fix $T > 0$ and $m \in \mathcal{N}$. Let $h = T/m$ and let $g \in C(\overline{\Omega})$. Define $u^m \in C([0,T) \times \overline{\Omega})$ by

$$u^m(t,x) = G_{t-lh} \circ \underbrace{G_h \circ \cdots \circ G_h}_{l \text{ times}} g(x) \tag{2.5}$$
$$(lh \leqq t < (l+1)h, l = 0, \dots, m-1, x \in \overline{\Omega}),$$

where G_0 is the identity operator on $C(\overline{\Omega})$.

Set

$$c_N = \frac{\int_0^{+\infty} f(r)r^N dr}{2(N-1)\int_0^{+\infty} f(r)r^{N-2}dr}.$$

Under the above situation, our main result is the following.

THEOREM 2.1. *Let $g \in C(\overline{\Omega})$. Let $\{u^m\}_{m\in\mathcal{N}}$ be a sequence of functions defined by* (2.5). *Then $u^m \to u$ locally uniformly on $[0,T) \times \overline{\Omega}$ as $m \to +\infty$. Here u is a unique viscosity solution of*

$$\begin{cases} u_t + c_N F(Du, D^2u) = 0 & \text{in} \quad (0,+\infty) \times \overline{\Omega}, \\ \dfrac{\partial u}{\partial n} = 0 & \text{in} \quad (0,+\infty) \times \partial\Omega, \\ u(0,x) = g(x) & x \in \overline{\Omega}. \end{cases} \tag{2.6}$$

Examples. Let $R > 0$ and $f(r) = 1$ for $0 \leqq r \leqq R$ and $= 0$ for $r > R$. Then $c_N = R^2/\{2(N+1)\}$. Let $f(r) = \exp(-r^2/4)$. Then $c_N = 1$.

In what follows we simply denote $c_N F$ by F and set $B_M(x,r) = \{y \in \mathcal{R}^M \mid |x-y| < r\}$.

3. Estimates for G_h. To show Theorem 2.1, the following estimates play a crucial role. In what follows we always assume (A.1)–(A.4).

Lemma 3.1. *Let $\varphi \in C^2(\overline{\Omega})$, $z \in \overline{\Omega}$, and $\varepsilon > 0$.*

(1) *Assume $z \in \Omega$ and $D\varphi(z) \neq 0$. Then there exists a $\delta > 0$ such that, for all $x \in B_N(z,\delta)$ and $h \in (0,\delta]$,*

$$G_h\varphi(x) \leqq \varphi(x) + (-F(D\varphi(z), D^2\varphi(z)) + \varepsilon)h, \tag{3.1}$$

$$G_h\varphi(x) \geqq \varphi(x) + (-F(D\varphi(z), D^2\varphi(z)) - \varepsilon)h. \tag{3.2}$$

(2) *Assume $z \in \partial\Omega$ and $(\partial\varphi/\partial n)(z) > 0$. Then there exists a $\delta > 0$ such that, for all $x \in B_N(z,\delta) \cap \overline{\Omega}$ and $h \in (0,\delta]$, (3.1) holds.*

(3) *Assume $z \in \partial\Omega$ and $(\partial\varphi/\partial n)(z) < 0$. Then there exists a $\delta > 0$ such that, for all $x \in B_N(z,\delta) \cap \overline{\Omega}$ and $h \in (0,\delta]$, (3.2) holds.*

Roughly speaking, we prove this lemma as follows. For any function g in $\mathcal{R}^N$, define $\widetilde{G}_h g$ by

$$\widetilde{G}_h g(x) = \sup\left\{\lambda \in \mathcal{R} \;\middle|\; \int_{\mathcal{R}^N} f_{\sqrt{h}}(y-x)\chi_{\{g \geqq \lambda\}}(y)dy \geqq \frac{1}{2}\int_{\mathcal{R}^N} f_{\sqrt{h}}(y-x)dy\right\}.$$

Then we compare $G_h\varphi(x)$ with $\widetilde{G}_h\varphi(x)$ and use the following lemma.

Lemma 3.2 (see [6, Theorem 3.1], [7, Lemma 3.1]). *Let $\varphi \in C^2(\mathcal{R}^N)$, $z \in \mathcal{R}^N$, and $\varepsilon > 0$. Assume $D\varphi(z) \neq 0$. Then there exists a $\delta > 0$ such that, for all $x \in B_N(z,\delta)$ and $h \in (0,\delta]$,*

$$\widetilde{G}_h\varphi(x) \leqq \varphi(x) + (-F(D\varphi(z), D^2\varphi(z)) + \varepsilon)h,$$
$$\widetilde{G}_h\varphi(x) \geqq \varphi(x) + (-F(D\varphi(z), D^2\varphi(z)) - \varepsilon)h.$$

As mentioned in the introduction, we divide our consideration into two subsections. In section 3.1, we consider the case where supp$f$ is compact. In section 3.2, we treat the noncompact case.

3.1. The case where supp*f* is compact. We prove Lemma 3.1 in the case where supp$f$ is compact. For simplicity, we assume supp$f \subset \overline{B_N(0,1)}$ and that there exists a $\rho \in (0,1]$ such that

$$f(y) > 0 \quad \text{in} \quad \{y \in B_N(0,1) \mid \operatorname{dist}(y, \partial B_N(0,1)) \leqq \rho\}. \tag{3.3}$$

Let $\varphi \in C^2(\overline{\Omega})$ and $\varepsilon > 0$. Assume $z \in \Omega$ and $D\varphi(z) \neq 0$. Then there exists a $\delta > 0$ such that, for all $x \in B_N(z,\delta) \subset \Omega$ and $h \in (0,\delta]$,

$$G_h\varphi(x) = \widetilde{G}_h\varphi(x).$$

Thus Lemma 3.2 yields Lemma 3.1, part (1). In the following we assume $z \in \partial\Omega$ and $(\partial\varphi/\partial n)(z) > 0$ and prove Lemma 3.1, part (2). Lemma 3.1, part (3) can be proved similarly.

Since $\partial\Omega$ is C^2 and $\varphi \in C^2(\overline{\Omega})$, there exists an $r_0 > 0$ such that $\varphi \in C^2(B_N(z,r_0))$, and we can take a $\psi \in C^2(\mathcal{R}^{N-1})$ for which

$$y_N - z_N = \psi(y' - z') \quad \text{for all } y = (y', y_N) \in B_N(z,r_0) \cap \partial\Omega,$$
$$D'\psi(0) = 0, \quad y_N - z_N > \psi(y' - z') \quad \text{for all } y \in B_N(z,r_0) \cap \Omega,$$

where $D' = (\partial/\partial x_1, \dots, \partial/\partial x_{N-1})$. Taking r_0 smaller, if necessary, we observe

$$D_N\varphi(y) \leqq -\gamma, \quad |D'\varphi(y)| \leqq K, \quad |D'\psi(y'-z')| \leqq \varepsilon \quad \text{for all } y \in B_N(z,r_0), \tag{3.4}$$

where $D_N = \partial/\partial x_N$ and γ, $K > 0$, are independent of ε and r_0.

Moreover, we prepare some lemmas.

LEMMA 3.3. *Let $r > 0$ and let a, $b \in \mathcal{R}^{N-1}$ and $c \in \mathcal{R}$. Define*

$$P^+ = \{y \in B_N(0,r) \mid y_N > \langle b, y'\rangle + c, y_N > \langle a, y'\rangle\},$$
$$P^- = \{y \in B_N(0,r) \mid y_N > \langle b, y'\rangle + c, y_N < \langle a, y'\rangle\}.$$

Assume that $\langle a, b\rangle + 1 \neq 0$.

(1) *There exists a $\theta \in (0,1)$ depending only and continuously on a and b such that if*

$$\{y \in B_N(0,r) \mid y_N = \langle b, y'\rangle + c, y_N = \langle a, y'\rangle\} \neq \emptyset,$$

then

$$\frac{|c|}{\sqrt{|b|^2+1}} \leqq \theta r.$$

(2) *Let θ be the same as above. For each $\theta_1 \in [\theta, 1)$, if*

$$\frac{|c|}{\sqrt{|b|^2+1}} \leqq \theta_1 r,$$

then

$$\int_{P^+} f_{\sqrt{h}}(y)dy \geqq \int_{P^-} f_{\sqrt{h}}(y)dy + \int_{S^-} f_{\sqrt{h}}(y)dy,$$

where $S^- = \{y \in B_N(0,r) \mid \langle y, n\rangle > \theta_1 r\}$ and $n = (b,-1)/\sqrt{|b|^2+1}$.

Remark 3.1. If we write $n = (b,-1)/\sqrt{|b|^2+1}$, then the equation $y_N = \langle b, y'\rangle + c$ implies

$$\langle n, y\rangle + \frac{c}{\sqrt{|b|^2+1}} = 0,$$

and thus the value $-c/\sqrt{|b|^2+1}$ is the distance of the plane $y_N = \langle b, y'\rangle + c$ from the origin in the direction n, where n is the normal unit vector to the plane $y_N = \langle b, y'\rangle + c$.

Proof of Lemma 3.3. (1) Fix $\xi = (\xi', \xi_N) \in \{y \in B_N(0,r) \mid y_N = \langle a, y'\rangle, y_N = \langle b, y'\rangle + c\}$, and set

$$n = \frac{(b,-1)}{\sqrt{|b|^2+1}}, \quad t = \langle \xi, n\rangle, \quad v = \xi - tn. \tag{3.5}$$

Since $\langle \xi, (a,-1)\rangle = 0$, we get

$$0 = \langle (v+tn), (a,-1)\rangle = \langle v, (a,-1)\rangle + t\frac{\langle a, b\rangle + 1}{\sqrt{|b|^2+1}}.$$

It is seen by $|v|^2 + t^2 = |\xi|^2 \leqq r^2$ that

$$t^2 = \frac{|b|^2+1}{(\langle a, b\rangle + 1)^2}\langle v, (a,-1)\rangle^2 \leqq \frac{|b|^2+1}{(\langle a, b\rangle + 1)^2}(r^2 - t^2)(|a|^2+1)$$

and therefore,

$$t^2\left(1+\frac{(|b|^2+1)(|a|^2+1)}{(\langle a,b\rangle+1)^2}\right)\leqq r^2\frac{(|b|^2+1)(|a|^2+1)}{(\langle a,b\rangle+1)^2}.$$

Setting

$$\theta=\left(\frac{(|b|^2+1)(|a|^2+1)}{(\langle a,b\rangle+1)^2+(|b|^2+1)(|a|^2+1)}\right)^{1/2}\left(\geqq\frac{1}{\sqrt{2}}\right),$$

we have

$$0<\theta<1,\quad |t|\leqq\theta r.$$

Moreover, θ is a continuous function of a and b.

(2) For simplicity, set $h=1$. Let $\theta_1\in[\theta,1)$ and assume

$$\frac{|c|}{\sqrt{|b|^2+1}}\leqq\theta_1 r.$$

Define $S^{\pm}\subset B_N(0,r)$ by

$$S^+=\{y\in B_N(0,r)\mid\langle y,n\rangle<-\theta_1 r\},\quad S^-=\{y\in B_N(0,r)\mid\langle y,n\rangle>\theta_1 r\}.$$

Then we have

$$\mathcal{L}^N(S^+)=\mathcal{L}^N(S^-)=\alpha r^N \tag{3.6}$$

for some $\alpha>0$ depending only and continuously on θ_1. Here $\mathcal{L}^N(A)$ is the N-dimensional Lebesgue measure of a set $A\subset\mathcal{R}^N$. From part (1), we see

$$\{y\in B_N(0,r)\mid y_N=\langle a,y'\rangle, y_N=\langle b,y'\rangle+c\}\cap(S^+\cup S^-)=\emptyset.$$

Define n and t as in (3.5). By an orthogonal change of variables, we may assume that $n=(0,-1)$, so that $b=0$ and $t=-c$.

Case 1. $c\geqq 0$ (cf. Figure 3.1).

Set

$$Q^+=\{y\in B_N(0,r)\mid y_N>\langle a,y'\rangle, c<y_N<\theta_1 r\},$$
$$Q^-=\{y\in B_N(0,r)\mid y_N<\langle a,y'\rangle, c<y_N<\theta_1 r\},$$
$$R_s^+=\partial B_N(0,s)\cap Q^+,\ R_s^-\ =\partial B_N(0,s)\cap Q^-.$$

Then, by geometry it is easily observed that

$$\mathcal{S}(R_s^+)\geqq\mathcal{S}(R_s^-)\quad\text{for all } s\in(c,\theta_1 r),$$

where $\mathcal{S}(R_s^{\pm})$ denotes the surface area of $R_s^{\pm}$. Since f satisfies (A.1) and (A.2) and S^+ is congruent to S^-, using the change of variables $s=|y|$, we can compute

$$\int_{Q^+}f(y)dy=\int_c^{\theta_1 r}f(s)\mathcal{S}(R_s^+)ds\geqq\int_c^{\theta_1 r}f(s)\mathcal{S}(R_s^-)ds=\int_{Q^-}f(y)dy.$$

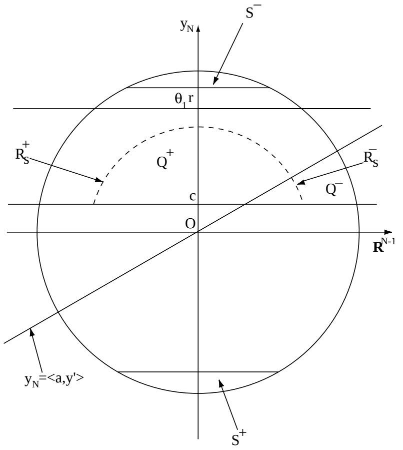


FIG. 3.1. *The case* $c \geqq 0$.

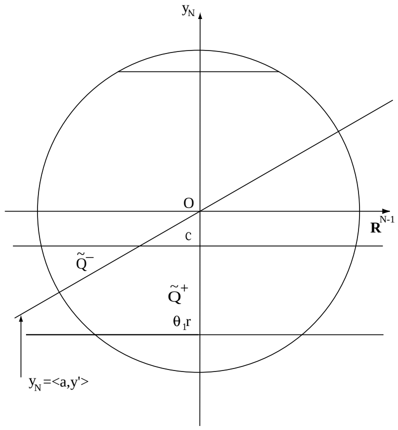


FIG. 3.2. *The case* $c < 0$.

Thus we obtain

$$\begin{aligned}\int_{P^+} f(y)dy &= \left(\int_{Q^+} + \int_{S^+}\right) f(y)dy \\ &\geqq \left(\int_{Q^-} + \int_{S^-}\right) f(y)dy = \left(\int_{P^-} + \int_{S^-}\right) f(y)dy.\end{aligned}$$

Case 2. $c < 0$ (cf. Figure 3.2).
Set

$$\begin{aligned}\widetilde{Q}^+ &= \{y \in B_N(0,r) \mid y_N > \langle a, y' \rangle, -\theta_1 r < y_N < c\}, \\ \widetilde{Q}^- &= \{y \in B_N(0,r) \mid y_N < \langle a, y' \rangle, -\theta_1 r < y_N < c\}.\end{aligned}$$

We easily see, as above, that

$$\begin{aligned}\int_{P^+} f(y)dy &= \frac{1}{2}\int_{B_N(0,1)} f(y)dy - \int_{\widetilde{Q}^+} f(y)dy \\ &\geqq \frac{1}{2}\int_{B_N(0,1)} f(y)dy - \int_{\widetilde{Q}^-} f(y)dy.\end{aligned}$$

Since

$$\int_{P^-} f(y)dy = \frac{1}{2}\int_{B_N(0,1)} f(y)dy - \int_{S^-} f(y)dy - \int_{\widetilde{Q}^-} f(y)dy,$$

we get

$$\int_{P^+} f(y)dy \geqq \int_{P^-} f(y)dy + \int_{S^-} f(y)dy. \qquad \square$$

Let $\widetilde{\lambda} = \widetilde{G}_h \varphi(x)$. Since $\varphi \in C^2(B_N(z, r_0))$, we can easily prove the following lemma.

Lemma 3.4. *There exist an $h_1 \in (0, r_0]$ and a $C_1 > 0$ independent of $x \in B_N(z, r_0)$ such that if $0 < h \leqq h_1$ and $x \in B_N(z, r_0)$, then $|\widetilde{\lambda} - \varphi(x)| \leqq C_1 h$.*

To continue, fix $C_2 > 0$ so that

$$\max_{B_{N-1}(0,r_0)} \|{D'}^2\psi\| \leqq C_2, \quad \max_{B_N(0,r_0)} \|D^2\varphi\| \leqq C_2.$$

Lemma 3.5. *There exist an $h_2 \in (0, r_0]$ and a $\theta_1 \in (0, 1)$ depending only on K, γ, C_1, and C_2 such that if $h \in (0, h_2)$ and*

$$\xi \in \{y \in B_N(x, \sqrt{h}) \mid \varphi(y) = \widetilde{\lambda}, y_N = \psi(y')\},$$

then $|\xi_N - x_N| \leqq \theta_1 \sqrt{h}$.

Proof. Using Lemma 3.4, we have

$$|\varphi(\xi) - \varphi(x)| = |\widetilde{\lambda} - \varphi(x)| \leqq C_1 h \quad (\text{for all } h \in (0, h_1)).$$

Since

$$|\varphi(\xi) - \varphi(x) - \langle D'\varphi(x), \xi' - x'\rangle - D_N\varphi(x)(\xi_N - x_N)| \leqq C_2 h, \quad D_N\varphi(x) \leqq -\gamma,$$

we get

$$\begin{aligned}\gamma|\xi_N - x_N| &\leqq |\varphi(\xi) - \varphi(x)| + |D'\varphi(x)||\xi' - x'| + C_2 h\\ &\leqq C_1 h + K\sqrt{h - |\xi_N - x_N|^2} + C_2 h.\end{aligned}$$

Hence, we see that

$$\gamma^2|\xi_N - x_N|^2 \leqq K^2(h - |\xi_N - x_N|^2) + (C_1 + C_2)^2 h^2 + 2(C_1 + C_2)h^{3/2}.$$

Thus

$$|\xi_N - x_N| \leqq \sqrt{\frac{K^2}{K^2 + \gamma^2} h + (C_1 + C_2)^2 h^2 + 2(C_1 + C_2)h^{3/2}}.$$

From this we complete the proof. □

Lemma 3.6. *There exists a $C_3 > 0$ depending only on γ, C_1, and C_2 such that if $y \in B_N(x, \sqrt{h})$ satisfies $\varphi(y) = \widetilde{\lambda}$, $h \in (0, h_1)$, and if $a = D'\varphi(x)/D_N\varphi(x)$, then $|(y_N - x_N) - \langle a, y' - x'\rangle| \leqq C_3 h$.*

Proof. Using Lemma 3.4, we observe that

$$C_1 h \geqq |\varphi(y) - \varphi(x)| = |\langle D'\varphi(x), y' - x'\rangle + D_N\varphi(x)(y_N - x_N) + O(h)|,$$

where $|O(h)| \leqq C_2 h$. Hence we have

$$|(y_N - x_N) - \langle a, y' - x'\rangle| \leqq (C_1 + C_2)h/\gamma. \qquad □$$

We are now in a position to prove Lemma 3.1, part (2). Set $\lambda = G_h\varphi(x)$ and $a \wedge b = \min\{a, b\}$.

Proof of Lemma 3.1, *part* (2). We show that there exists an $h_0 > 0$ such that $\lambda \leqq \widetilde{\lambda}$ for all $x \in B_N(z, r_0) \cap \overline{\Omega}$ and $h \in (0, h_0)$. Then we use Lemma 3.2, part (2) to obtain our desired result.

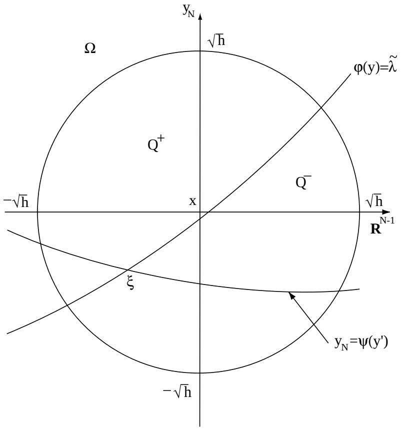


FIG. 3.3. *Case* 1.

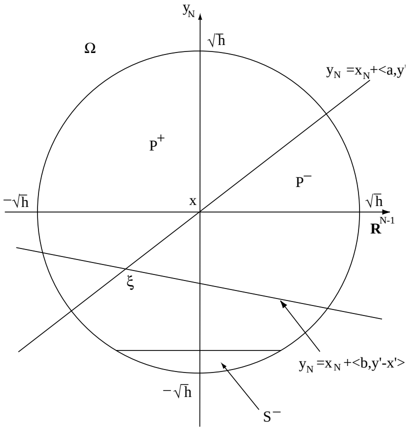


FIG. 3.4. *The sets* $P^{\pm}$ *and* S.

Define

$$Q^+ = \{y \in B_N(x, \sqrt{h}) \mid \varphi(y) < \widetilde{\lambda}, y_N > \psi(y')\},$$
$$Q^- = \{y \in B_N(x, \sqrt{h}) \mid \varphi(y) > \widetilde{\lambda}, y_N > \psi(y')\},$$
$$\widetilde{Q}^+ = \{y \in B_N(x, \sqrt{h}) \mid \varphi(y) < \widetilde{\lambda}\}, \widetilde{Q}^- = \{y \in B_N(x, \sqrt{h}) \mid \varphi(y) > \widetilde{\lambda}\}.$$

Note that $Q^{\pm} = \widetilde{Q}^{\pm} \cap \Omega$ for small $h > 0$. Here and in what follows we denote $\psi(y' - z') - z_N$ by $\psi(y')$ for simplicity. We observe

$$\int_{B_N(x,\sqrt{h})} f_{\sqrt{h}}(y-x)\chi_{\{\varphi \geqq \widetilde{\lambda}\}}(y)dy = \frac{1}{2}\int_{B_N(x,\sqrt{h})} f_{\sqrt{h}}(y-x)dy,$$

and thus

$$\int_{\widetilde{Q}^+} f_{\sqrt{h}}(y-x)dy = \int_{\widetilde{Q}^-} f_{\sqrt{h}}(y-x)dy. \tag{3.7}$$

Case 1. $\{y \in B_N(x, \sqrt{h}) \mid \varphi(y) = \widetilde{\lambda}, y_N = \psi(y')\} \neq \emptyset$ (cf. Figure 3.3).

Let $h_0 \leqq h_1 \wedge h_2$ and assume that $h \in (0, h_0)$. Fix $\xi \in \{y \in B_N(x, \sqrt{h}) \mid \varphi(y) = \widetilde{\lambda}, y_N = \psi(y')\}$. For $y' \in B_{N-1}(x', \sqrt{h})$,

$$\psi(y') = \psi(\xi') + \langle D'\psi(\xi'), y' - \xi'\rangle + O(h),$$

where $|O(h)| \leqq C_2 h$. Hence, if $y \in B_N(x, \sqrt{h})$ satisfies $y_N = \psi(y')$, then

$$|(y_N - \xi_N) - \langle D'\psi(\xi'), y' - \xi'\rangle| \leqq C_2 h. \tag{3.8}$$

Set

$$a(x) = \frac{D'\varphi(x)}{D_N\varphi(x)}, \quad b(\xi') = D'\psi(\xi'), \quad c(x;\xi) = \xi_N - x_N - \langle b(\xi'), \xi' - x'\rangle$$

for $x \in B_N(z, r_0)$ and $\xi \in \{y \in B_N(x, \sqrt{h}) \mid \varphi(y) = \widetilde{\lambda}, y_N = \psi(y')\}$. We observe that if $h \in (0, h_0)$, then

$$|\langle a(x), b(\xi')\rangle + 1| \geqq 1 - \frac{\varepsilon K}{\gamma}, \quad \frac{|c(x;\xi)|}{\sqrt{|b(\xi')|^2 + 1}} \leqq (\theta_1 + \varepsilon)\sqrt{h},$$

where θ_1 is from Lemma 3.5. We have used $|b(\xi')| = |D'\psi(\xi')| \leqq \varepsilon$ in (3.4). Letting $\theta_2 = (1+\theta_1)/2$ and $\varepsilon_0 = \min\{\gamma/K, \theta_2 - \theta_1\}/2$, for any $\varepsilon \in (0, \varepsilon_0)$, we get

$$|\langle a(x), b(\xi')\rangle + 1| \geqq \frac{1}{2}, \quad \frac{|c(x;\xi)|}{\sqrt{|b(\xi')|^2+1}} \leqq \theta_2\sqrt{h}.$$

For any $x \in B_N(z, r_0)$ and $\xi \in \{y \in B_N(x, \sqrt{h}) \mid \varphi(y) = \widetilde{\lambda}, y_N = \psi(y')\}$, put

$$a = a(x), \quad b = b(\xi'), \quad c = c(x;\xi), \quad \widetilde{n} = \frac{(b,-1)}{\sqrt{|b|^2+1}}.$$

In view of Lemma 3.3, part (2), we have

$$\int_{P^+} f_{\sqrt{h}}(y-x)dy \geqq \int_{P^-} f_{\sqrt{h}}(y-x)dy + \int_{S^-} f_{\sqrt{h}}(y-x)dy,$$

where

$$\begin{aligned}
P^+ &= \{y \in B_N(x,\sqrt{h}) \mid y_N > x_N + \langle b, y'-x'\rangle + c, y_N > x_N + \langle a, y'-x'\rangle\},\\
P^- &= \{y \in B_N(x,\sqrt{h}) \mid y_N > x_N + \langle b, y'-x'\rangle + c, y_N < x_N + \langle a, y'-x'\rangle\},\\
S^- &= \{y \in B_N(x,\sqrt{h}) \mid \langle y-x, \widetilde{n}\rangle > \theta_1\sqrt{h}\}
\end{aligned}$$

(cf. Figure 3.4). Note that the estimate of $\int_{S^-} f_{\sqrt{h}}(y-x)dy$ is independent of $x \in B_N(z, r_0)$, $\varepsilon \in (0, \varepsilon_0)$, and $h > 0$ because of (3.3) and (3.6) with $r = \sqrt{h}$. We easily see, by Lemma 3.6 and (3.8), that

$$\begin{aligned}
&\{y \in B_N(x,\sqrt{h}) \mid \varphi(y) = \widetilde{\lambda}\}\\
&\quad \subset \{y \in B_N(x,\sqrt{h}) \mid |y_N - x_N - \langle a, y'-x'\rangle| \leqq C_3 h\} \equiv T_1,\\
&B_N(x,\sqrt{h}) \cap \partial\Omega\\
&\quad \subset \{y \in B_N(x,\sqrt{h}) \mid |y_N - x_N - \langle b, y'-x'\rangle - c| \leqq C_2 h\} \equiv T_2.
\end{aligned}$$

Moreover, we compute

$$\begin{aligned}
\int_{Q^+} f_{\sqrt{h}}(y-x)dy = \int_{\widetilde{Q}^+\cap\Omega} f_{\sqrt{h}}(y-x)dy \geqq \int_{P^+} f_{\sqrt{h}}(y-x)dy - \int_{T_1} f_{\sqrt{h}}(y-x)dy,\\
\int_{Q^-} f_{\sqrt{h}}(y-x)dy = \int_{\widetilde{Q}^-\cap\Omega} f_{\sqrt{h}}(y-x)dy \leqq \int_{P^-} f_{\sqrt{h}}(y-x)dy + \int_{T_2} f_{\sqrt{h}}(y-x)dy,
\end{aligned}$$

and therefore we have

$$\int_{Q^+} f_{\sqrt{h}}(y-x)dy \geqq \int_{Q^-} f_{\sqrt{h}}(y-x)dy + \int_{S^-} f_{\sqrt{h}}(y-x)dy - \left(\int_{T_1} + \int_{T_2}\right) f_{\sqrt{h}}(y-x)dy.$$

By Lemma 3.6 we have $\mathcal{L}^N(T_1) \leqq C_4 h^{(N+1)/2}$ for some $C_4 > 0$. The smoothness of $\partial\Omega$ yields $\mathcal{L}^N(T_2) \leqq C_5 h^{(N+1)/2}$ for some $C_5 > 0$. Changing the variables $(y-x)/\sqrt{h} \to y$, we observe

$$\begin{aligned}
&\left(\int_{T_1} + \int_{T_2}\right) f_{\sqrt{h}}(y-x)dy = \left(\int_{T_1(h,x)} + \int_{T_2(h,x)}\right) f(y)dy,\\
&\mathcal{L}^N(T_1(h,x)) \leqq C_4\sqrt{h}, \quad \mathcal{L}^N(T_2(h,x)) \leqq C_5\sqrt{h},\\
&(T_i(h,x) = \{(y-x)/\sqrt{h} \mid y \in T_i\}, i = 1, 2).
\end{aligned}$$

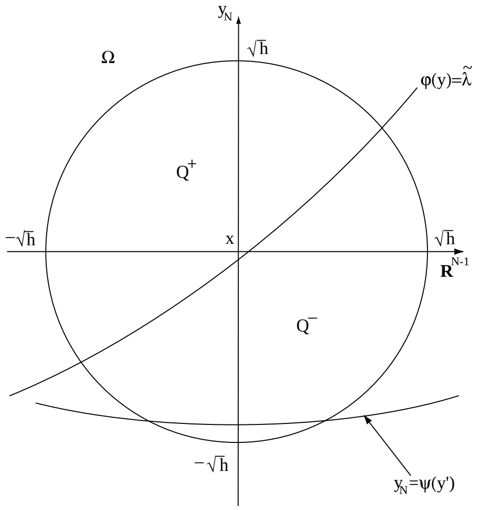


FIG. 3.5. *Case* 2.

Hence, taking $h_0 > 0$ smaller, if necessary, we obtain $\lambda \leqq \widetilde{\lambda}$ for any $x \in B_N(z, r_0)$ and $h \in (0, h_0)$.

Case 2. $\{y \in B_N(x, \sqrt{h}) \mid \varphi(y) = \widetilde{\lambda}\} \subset \Omega$ (cf. Figure 3.5).

In this case we have

$$\widetilde{Q}^+ = Q^+, \quad Q^- \subset \widetilde{Q}^-,$$

and hence we get $\lambda \leqq \widetilde{\lambda}$ by (3.7).

Case 3. $\{y \in B_N(x, \sqrt{h}) \mid \varphi(y) = \widetilde{\lambda}\} \subset \Omega^c$.

In this case we have $Q^- = \emptyset$, and hence $\lambda \leqq \widetilde{\lambda}$ is trivial.

Therefore, we have completed the proof. □

Remark 3.2. The assumptions we have actually used are only (A.1) and $\int_{\mathcal{R}^N} f(y)dy < +\infty$ because we have assumed that supp$f$ is compact. The other assumptions play important roles in the next subsection.

3.2. The case where supp*f* is not compact. By (A.2) there exist a $\rho_1 > 0$, $\{R(\rho)\}_{0<\rho<\rho_1}$ and an $\omega \in C(\mathcal{R}^+; \mathcal{R}^+)$ such that

$$\begin{aligned} &\int_{B_N(0,R(\sqrt{h}))^c} f(y)dy \leqq \sqrt{h}\omega(R(\sqrt{h})) \quad (h \in (0, \rho_1)), \\ &\omega(R) \to 0 \ (R \to +\infty), \quad R(\rho) \to +\infty, \quad \rho R(\rho)^2 \to 0 \quad (\rho \to 0) \end{aligned} \tag{3.9}$$

(cf. Ishii, Pires, and Souganidis [7]).

Moreover, (A.4) implies that, for each $\varepsilon > 0$, there exists an $\rho_2 = \rho_2(\varepsilon) > 0$ such that for any $g(\xi) = \langle A\xi, \xi \rangle + a$ $(A \in \mathcal{S}^{N-1},\ a \in \mathcal{R})$,

$$\sup_{0<r<\rho_2} \left| \int_{B_{N-1}(0,R(\rho_2))} f(y', rg(y'))g(y')dy' - \int_{\mathcal{R}^{N-1}} f(y', 0)g(y')dy' \right| < \varepsilon. \tag{3.10}$$

For $r > 0$, define

$$\widetilde{S}_h^r\varphi(x) = \int_{B_N(x,\sqrt{h}r)} f_{\sqrt{h}}(x-y)\varphi(y)dy,$$
$$\widetilde{G}_h^r\varphi(x) = \sup\left\{\lambda \in \mathcal{R} \;\middle|\; \widetilde{S}_h^r\chi_{\{\varphi \geqq \lambda\}}(x) \geqq \frac{1}{2}\int_{B_N(x,\sqrt{h}r)} f_{\sqrt{h}}(y-x)dy\right\},$$

and let $\widetilde{\mu}(r) = \widetilde{G}_h^r\varphi(x)$.

Fix $\varphi \in C^2(\overline{\Omega})$ and $\varepsilon > 0$. Assume $z \in \overline{\Omega}$ and $D\varphi(z) \neq 0$. Then there exists an $r_1 > 0$ such that $\varphi \in C^2(B_N(z,r_1))$ and $D\varphi \neq 0$ in $B_N(z,r_1)$. We can easily verify

$$\int_{B_N(x,\sqrt{h}r)} f_{\sqrt{h}}(y-x)\chi_{\{\varphi \geqq \widetilde{\mu}(r)\}}(y)dy = \frac{1}{2}\int_{B_N(x,\sqrt{h}r)} f_{\sqrt{h}}(y-x)dy \tag{3.11}$$

for all $r > 0$ satisfying $\sqrt{h}r < r_1$.

For any $h \in (0,\rho_1)$, set $\widetilde{\mu} = \widetilde{\mu}(R(\sqrt{h}))$. As to the estimates for $\widetilde{\mu}$ and $\widetilde{\lambda}$, we have the following lemmas.

LEMMA 3.7. *There exist an $h_1 \in (0,r_1]$ and a $C_1 > 0$ independent of $x \in B_N(z,r_1) \cap \overline{\Omega}$ such that if $x \in B_N(z,r_1)$ and $h \in (0,h_1)$, then $|\widetilde{\lambda} - \varphi(x)|$, $|\widetilde{\mu} - \varphi(x)| \leqq C_1 h$.*

This lemma can be shown easily, so we omit the proof.

LEMMA 3.8. *There exist an $\varepsilon_1 > 0$ and a $C_2 > 0$ such that, for any $\varepsilon \in (0,\varepsilon_1)$, there exists an $h_2 > 0$ such that $|\widetilde{\lambda} - \widetilde{\mu}| \leqq C_2\varepsilon h$ for all $h \in (0,h_2)$.*

Proof. It is easily seen by the change of variables $(y-x)/\sqrt{h} \to y$, (3.11), and (3.9) that

(3.12)
$$\begin{aligned}
&\int_{\mathcal{R}^N} f_{\sqrt{h}}(y-x)\chi_{\{\varphi \geqq \widetilde{\mu} - C_2\varepsilon h\}}(y)dy \\
&\quad = \int_{\mathcal{R}^N} f(y)\chi_{\{\varphi \geqq \widetilde{\mu} - C_2\varepsilon h\}}(x - \sqrt{h}y)dy \\
&\quad \geqq \int_{B_N(0,\sqrt{h}R(\sqrt{h}))} f(y)\chi_{\{\varphi \geqq \widetilde{\mu}\}}(x - \sqrt{h}y)dy \\
&\qquad + \int_{B_N(0,\sqrt{h}R(\sqrt{h}))} f(y)\chi_{\{\mu > \varphi \geqq \widetilde{\mu} - C_2\varepsilon h\}}(x - \sqrt{h}y)dy \\
&\quad = \frac{1}{2}\int_{B_N(0,\sqrt{h}R(\sqrt{h}))} f(y)dy + \int_{B_N(0,\sqrt{h}R(\sqrt{h}))} f(y)\chi_{\{\mu > \varphi \geqq \widetilde{\mu} - C_2\varepsilon h\}}(x - \sqrt{h}y)dy \\
&\quad \geqq \frac{1}{2}\int_{\mathcal{R}^N} f(y)dy - \sqrt{h}\omega(R(\sqrt{h})) \\
&\qquad + \int_{B_N(0,\sqrt{h}R(\sqrt{h}))} f(y)\chi_{\{\mu > \varphi \geqq \widetilde{\mu} - C_2\varepsilon h\}}(x - \sqrt{h}y)dy
\end{aligned}$$

for all $h \in (0,\rho_1 \wedge \rho_2)$.

We take a continuous family $\{U(x)\}_{x \in B_N(z,r_1)} \subset \mathcal{O}(N)$ satisfying

$$U(x)\left(\frac{D\varphi(x)}{|D\varphi(x)|}\right) = e_N \quad \text{for all } x \in B_N(z,r_1).$$

By the same calculations as in [7], we observe that there exists a $\delta_1 > 0$ such that if

$$\varphi(x - \sqrt{h}U^*(x)y) \geqq \widetilde{\mu} \text{ (resp., } \varphi(x - \sqrt{h}U^*(x)y) \leqq \widetilde{\mu}),$$

then

$$y_N \leqq \frac{\sqrt{h}}{|D\varphi(z)|}\left(\frac{\varphi(x)-\widetilde{\mu}}{h} + \varepsilon + \frac{1}{2}\langle P^*U(z)D^2\varphi(z)PU^*(z)Py', y'\rangle + \varepsilon|y'|^2\right)$$

$$\Bigg(\text{resp.,}$$

$$y_N \geqq \frac{\sqrt{h}}{|D\varphi(z)|}\left(\frac{\varphi(x)-\widetilde{\mu}}{h} + \varepsilon + \frac{1}{2}\langle P^*U(z)D^2\varphi(z)PU^*(z)Py', y'\rangle + \varepsilon|y'|^2\right)\Bigg)$$

for $\sqrt{h}y \in B_N(0,\delta_1)$ and $x \in B_N(z,r_1)$. Here P denotes the $N \times (N-1)$ matrix, whose (i,j)th entries are 1 if $i = j$ and 0 if $i \neq j$. A similar inequality to the one above holds for $\varphi(x - \sqrt{h}U^*(x)y) \leqq (\geqq)\widetilde{\mu} - C_2\varepsilon h$. Here we take δ_1 and r_1 smaller, if necessary.

For simplicity, set $U(z) = I$ and $|D\varphi(z)| = 1$. Then we see

$$\begin{aligned}
&\int_{B_N(0,R(\sqrt{h}))} f(y)\chi_{\{\widetilde{\mu}>\varphi\geqq\widetilde{\mu}-C_2\varepsilon h\}}(x-\sqrt{h}y)dy \\
&\geqq \int_{B_{N-1}(0,R(\rho_2))}\int_{\sqrt{h}(\psi_1(y')+\varepsilon(1+|y'|^2))}^{\sqrt{h}(\psi_1(y')-\varepsilon(1+|y'|^2)+C_2\varepsilon)} f(y',y_N)dy_Ndy' \\
&\equiv J,
\end{aligned}$$

where $\psi_1(y') = \langle P^*D^2\varphi(z)Py', y'\rangle/2$.

By (3.10) we observe

$$J \geqq \varepsilon\sqrt{h}\left(-2 - 2\int_{\mathcal{R}^{N-1}} f(y',0)|y'|^2dy' + (C_2 - 2)\int_{\mathcal{R}^{N-1}} f(y',0)dy'\right).$$

Thus, taking a $C_2 > 0$ large and an $h_2 \in (0, \rho_1 \wedge \rho_2)$ small, we have

$$\int_{B_N(0,R(\sqrt{h}))} f(y)\chi_{\{\widetilde{\mu}>\varphi\geqq\widetilde{\mu}-C_2\varepsilon h\}}(x-\sqrt{h}y)dy \geqq \varepsilon\sqrt{h} \geqq \sqrt{h}\omega(R(\sqrt{h}))$$

for all $h \in (0, h_2)$. Thus, from (3.12), we conclude $\widetilde{\lambda} \geqq \widetilde{\mu} - C_2\varepsilon h$. Since we can show $\widetilde{\lambda} \leqq \widetilde{\mu} + C_2\varepsilon h$ by a similar argument to the above, we have completed the proof. □

Consider the case $z \in \Omega$. Then there exists a $\delta_2 > 0$ such that

$$B_N(z,\delta_2) \subset\subset \Omega,\ D\varphi(x) \neq 0 \quad (\text{for all } x \in B_N(z,\delta_2)).$$

Replacing ε_1, $h_2 > 0$ with smaller ones, if necessary, we see that $|\widetilde{\lambda} - \widetilde{\mu}|$, $|\lambda - \widetilde{\mu}| \leqq C_2\varepsilon h$, for all $\varepsilon \in (0,\varepsilon_1)$ and $h \in (0,h_2)$. Hence we have $|\lambda - \widetilde{\lambda}| \leqq 2C_4\varepsilon h$, and using Lemma 3.2 and this, we can prove Lemma 3.1, part (1). Therefore, we assume $z \in \partial\Omega$ and $(\partial\varphi/\partial n)(z) > 0$ and show Lemma 3.1, part (2). Lemma 3.1, part (3) can be proved similarly. We use the same notations r_0, ψ, D', ..., as those in the previous subsection.

It is easily seen that Lemma 3.3 holds. Moreover, we have the following lemmas by similar arguments to the proofs of Lemmas 3.5 and 3.6. Let $C_3 > 0$ satisfy $\max_{B_{N-1}(0,r_0)} \|D'^2\psi\| \leqq C_3$ and $\max_{B_N(0,r_0)} \|D^2\varphi\| \leqq C_3$.

Lemma 3.9. *There exist an $h_3 > 0$ and a $\theta_1 \in (0,1)$ depending only on γ, K, C_2, and C_3 such that if $h \in (0, h_3)$ and*

$$\xi \in \{y \in B_N(x, \sqrt{h}R(\sqrt{h})) \mid \varphi(y) = \widetilde{\mu}, y_N = \psi(y')\},$$

then $|\xi_N - x_N| \leqq \theta_1 \sqrt{h}R(\sqrt{h})$.

Lemma 3.10. *There is a $C_4 > 0$ depending only on γ, C_1, and C_3 such that if $\widetilde{\mu} = \widetilde{G}_h\varphi(x)$, $y \in B_N(x, \sqrt{h}R(\sqrt{h}))$, satisfies $\varphi(y) = \widetilde{\mu}$, $h \in (0, h_1)$, and $a = D'\varphi(x)/D_N\varphi(x)$, then $|(y_N - x_N) - \langle a, y' - x'\rangle| \leqq C_4 hR(\sqrt{h})^2$.*

To estimate the integration of $f_{\sqrt{h}}(\cdot - x)$ in Ω^c, we need the following lemmas.

Lemma 3.11. *There exists a $C_5 > 0$ depending only on γ, C_1, and C_3 for which*

$$\sup_{\substack{y \in B_N(x, \sqrt{h}R(\sqrt{h})) \cap \{\varphi(y) = \widetilde{\mu}\} \\ x \in B_N(z, r_0)}} \left\langle \frac{D\varphi(x)}{|D\varphi(x)|}, y - x \right\rangle \leqq C_5 hR(\sqrt{h})^2 \quad \textit{for all } h \in (0, h_1).$$

This lemma can be shown easily, so we omit the proof.

Lemma 3.12. *Let $0 < s < r$. Let $a \in \mathcal{R}^N$ satisfy $\langle a, e_N\rangle < 0$ and let $b \in \mathcal{R}^N$ be the projection of a to $\{y \in \mathcal{R}^N \mid \langle y, e_N\rangle = 0\}$. Assume $b \neq 0$. For $\alpha > 0$, define*

$$\begin{aligned} Q &= \{y \in B_N(0, r) \mid \langle a, y\rangle \leqq 0, |y_N - s| \leqq \alpha\}, \\ R &= \{y \in B_N(0, r) \cap B_N(0, s + \alpha)^c \mid \langle a, y\rangle \geqq 0, \langle b, y\rangle \leqq 0\}. \end{aligned}$$

Then there exists an $\alpha_0 > 0$ such that, for any $\alpha \in (0, \alpha_0)$,

$$\int_Q f(y)dy \leqq \int_R f(y)dy.$$

Proof. Take $\alpha_0 > 0$ so small that, for $\alpha \in (0, \alpha_0)$,

$$\{y \in Q \mid \langle y, e_N\rangle \geqq -\alpha\} = \emptyset, \quad \operatorname{dist}(0, Q \cup R) \geqq \frac{1}{2}s.$$

We introduce the polar coordinate system

$$y = \frac{t}{\sqrt{1 + \eta^2}}(\xi, \eta) \quad (t \geqq 0, \eta \in \mathcal{R}, \xi \in S^{N-2}).$$

Take $\xi_0 \in S^{N-2}$ and define $\eta_0 < 0$ by $\langle a, (\xi_0, \eta_0)\rangle = 0$. Let $s_1 \in (s + \alpha, r)$ such that $\frac{s_1\eta_0}{\sqrt{1+\eta_0^2}} - s = \alpha$. Moreover, define η_1, $\eta_2 < 0$, $(\eta_0 < \eta_2 < \eta_1 < 0)$, by

$$\frac{t\eta_1}{\sqrt{1 + \eta_1^2}} - s = \alpha, \quad \frac{t\eta_2}{\sqrt{1 + \eta_2^2}} - s = -\alpha$$

for each $t \in [s_1, r]$. By the choice of a and α_0, we get $\sup_{t \in [s_1, r]} |\eta_1 - \eta_2| \leqq C_6\alpha$ for $C_6 > 0$ independent of $\alpha \in (0, \alpha_0)$. Therefore,

$$\begin{aligned} \int_Q f(y)dy &= \int_{s_1}^{r} \int_{S_1^{N-2}} \int_{\eta_2}^{\eta_1} f(t)t^{N-1} \frac{1}{(1 + \eta^2)^{N/2}} d\eta d\xi dt \\ &\leqq C_6\alpha \int_{s_1}^{r} \int_{S_1^{N-2}} f(t)t^{N-1} d\xi dt, \end{aligned} \tag{3.13}$$

where $S_1^{N-2} = \{y \in S^{N-2} \mid \langle a, y\rangle \leqq 0\}$.

On the other hand, it is easily observed that

$$\int_R f(y)dy = \int_{s+\alpha}^{r} \int_{S_1^{N-2}} \int_{-\infty}^{\eta_0} f(t)t^{N-1} \frac{1}{(1+\eta^2)^{N/2}} d\eta d\xi dt.$$

Thus, taking $\alpha_0 > 0$ smaller, if necessary, by (3.13) and this equality, we obtain our desired result for all $\alpha \in (0, \alpha_0)$. □

LEMMA 3.13. *Let $0 < s < r$ and let a, $b \in \mathcal{R}^N$, be the same as in the above lemma. Define*

$$\widetilde{Q} = \{y \in B_N(0,r) \mid \langle a, y\rangle \leqq 0, y_N \leqq -(s+\alpha)\},$$
$$\widetilde{R} = \{y \in B_N(0,r) \mid \langle a, y\rangle \geqq 0, \langle b, y\rangle \geqq 0, y_N \leqq -(s+\alpha)\}.$$

Then

$$\int_{\widetilde{Q}} f(y)dy \leqq \int_{\widetilde{R}} f(y)dy$$

for all $\alpha > 0$.

Proof. Let $\widetilde{Q}_t = \partial B_N(0,t) \cap \widetilde{Q}$ and $\widetilde{R}_t = \partial B_N(0,t) \cap \widetilde{R}$. Then it is easily seen that $\mathcal{S}(\widetilde{Q}_t) \leqq \mathcal{S}(\widetilde{R}_t)$ for all $t \in [s+\alpha, r]$. Here $\mathcal{S}(A)$ denotes the area for a surface $A \subset \mathcal{R}^N$. Thus, changing the variables $t = |y|$, we have

$$\int_{\widetilde{Q}} f(y)dy = \int_{s+\alpha}^{r} f(t)\mathcal{S}(\widetilde{Q}_t)dt \leqq \int_{s+\alpha}^{r} f(t)\mathcal{S}(\widetilde{R}_t)dt = \int_{\widetilde{R}} f(y)dy. \qquad \square$$

We are now in a position to prove Lemma 3.1, part (2).

Proof of Lemma 3.1, *part* (2). It suffices to show that there exist a $C > 0$ and an $\varepsilon_0 > 0$ such that, for any $\varepsilon \in (0, \varepsilon_0)$, there exists an $h_0 > 0$ such that

$$\lambda \leqq \widetilde{\lambda} + C\varepsilon h \quad (\text{for all } x \in B_N(z, r_0) \cap \overline{\Omega}, h \in (0, h_0)).$$

Once we have the above inequality, by using Lemma 3.2 and this inequality, we obtain our desired result.

Case 1. $\{y \in B_N(x, \sqrt{h}R(\sqrt{h})) \mid \varphi(y) = \widetilde{\mu}\} \cap \partial\Omega \neq \emptyset$.

We may assume $R(\sqrt{h}) \geqq R_1 = R(\rho_1/2)$ for all $h \in (0, \rho_1^2/4)$.

Subcase 1-1. $\mathrm{dist}(x, \partial\Omega) \leqq \sqrt{h}R_1$.

Lemma 3.3 holds with $r = \sqrt{h}R(\sqrt{h})$. Let θ_1 and S^- be the same as in Lemma 3.3. By a similar argument in Case 1 in section 3.1, there exists an $\varepsilon_1 > 0$ such that

$$\int_{S^-} f_{\sqrt{h}}(y-x)dy \geqq \int_{B_N(x, \sqrt{h_1/2}R_1 \cap \{\langle y,n\rangle \geqq \theta_1 R(\rho_1/2)\}} f(y)dy > 0$$

for all $h \in (0, \rho_1^2/4)$, $x \in B_N(z, r_0)$, and $\varepsilon \in (0, \varepsilon_1)$. Thus we have

$$\begin{aligned} &\int_\Omega f_{\sqrt{h}}(y-x)\chi_{\{\varphi \geqq \widetilde{\mu}\}}(y)dy - \sqrt{h}\omega(R(\sqrt{h})) + \int_{S^-} f_{\sqrt{h}}(y-x)dy \\ &\quad \leqq \int_{\Omega \cap B_N(x, \sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)\chi_{\{\varphi \geqq \widetilde{\mu}\}}(y)dy + \int_{S^-} f_{\sqrt{h}}(y-x)dy = *. \end{aligned}$$

A similar argument in the previous subsection yields

$$\begin{aligned} * &\leqq \int_{\Omega \cap B_N(x, \sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)\chi_{\{\varphi < \widetilde{\mu}\}}(y)dy + \left(\int_{T_1} + \int_{T_2}\right) f_{\sqrt{h}}(y-x)dy \\ &\leqq \int_\Omega f_{\sqrt{h}}(y-x)\chi_{\{\varphi < \widetilde{\mu}\}}(y)dy + \left(\int_{T_1} + \int_{T_2}\right) f_{\sqrt{h}}(y-x)dy, \end{aligned}$$

where T_1 and T_2 are defined by

$$T_1 = \{y \in B_N(x, \sqrt{h}R(\sqrt{h})) \mid |y_N - x_N - \langle a, y' - x' \rangle| \leqq C_3 h R(\sqrt{h})^2\},$$
$$T_2 = \{y \in B_N(x, \sqrt{h}R(\sqrt{h})) \mid |y_N - x_N - \langle b, y' - x' \rangle - c| \leqq C_2 h R(\sqrt{h})^2\},$$
$$\left(a = \frac{D'\varphi(x)}{D_N\varphi(x)}, b = D'\psi(\xi'), c = \xi_N - x_N - \langle b, \xi' - x' \rangle \right).$$

By similar calculations to the proof of Lemma 3.8, we get

$$\int_{T_1} f_{\sqrt{h}}(y - x)dy \leqq C_6 \sqrt{h} R(\sqrt{h})^2$$

for some $C_6 > 0$. On the other hand, we easily see that

$$\int_{T_2 \cap B_N(x, \sqrt{h}R_1)} f_h(y - x)dy \leqq \widetilde{\omega}(h)$$

for some $\widetilde{\omega} \in C(\mathcal{R}^+; \mathcal{R}^+)$ with $\widetilde{\omega}(0) = 0$. Using a similar argument to the proof of Lemma 3.12, we have

$$\int_{T_2 \cap B_N(x, \sqrt{h}R_1)^c} f_h(y - x)dy \leqq C_7 \sqrt{h} R(\sqrt{h})^2$$

for some $C_7 > 0$. Note that C_6, C_7 are independent of ε, $h > 0$, and $x \in B_N(z, r_0)$ and that $\widetilde{\omega}$ is independent of $\varepsilon > 0$ and $x \in B_N(z, r_0)$. Therefore, we can take $h_4 \in (0, \rho_1^2/4)$ to satisfy

$$\int_{S^-} f_{\sqrt{h}}(y - x)dy \geqq \sqrt{h}\omega(R(\sqrt{h})) + \left(\int_{T_1} + \int_{T_2} \right) f_{\sqrt{h}}(y - x)dy$$

for all $h \in (0, h_4)$. Then we have $\lambda \leqq \widetilde{\mu} \leqq \widetilde{\lambda} + C_2 \varepsilon h$ for all $h \in (0, h_4)$.

Subcase 1-2. $\operatorname{dist}(x, \partial\Omega) > \sqrt{h}R_1$.

Lemma 3.11 yields that, for some $h_5 = h_5(R_1, \gamma) > 0$,

$$\{y \in \Omega^c \mid \varphi(y) \leqq \widetilde{\mu}\} \subset \left\{ y \in \mathcal{R}^N \;\middle|\; \left\langle \frac{D\varphi(x)}{|D\varphi(x)|}, y - x \right\rangle \leqq 1 \right\}$$

for all $x \in B_N(z, r_0) \cap \overline{\Omega}$ and $h \in (0, h_5)$). Then, for each $x \in B_N(z, r_0) \cap \overline{\Omega}$, we can choose a unit vector $a = a(x, R_1) \in \mathcal{R}^N$ such that

$$\langle a, n(z_x) \rangle \geqq \frac{1}{2} \inf_{y \in B_N(z, r_0) \cap \overline{\Omega}} \left\langle \frac{D\varphi(y)}{|D\varphi(y)|}, n(z_y) \right\rangle,$$
$$\{y \in \Omega^c \mid \varphi(y) \leqq \widetilde{\mu}\} \subset \{y \in \mathcal{R}^N \mid \langle a, y - x \rangle \leqq 0\},$$

where $z_x \in \partial\Omega$ is a unique point satisfying $\operatorname{dist}(x, \partial\Omega) = |x - z_x|$. On the other hand, we observe that there exists an $h_6 > 0$ such that

$$\partial\Omega \cap B_N(x, \sqrt{h}R(\sqrt{h})) \subset \{y \in B_N(x, \sqrt{h}R(\sqrt{h})) \mid \langle a, y - x \rangle \leqq 0, |y_N - z_{x,N}| \leqq C_8 h R(\sqrt{h})^2\}$$

for all $h \in (0, h_6)$, and $C_8 > 0$ is independent of $x \in B_N(z, r_0)$, ε, $h > 0$, and $x \in B_N(z, r_0)$. Define

$$\begin{aligned}
Q &= \{y \in B_N(x, \sqrt{h}R(\sqrt{h})) \mid \langle a, y-x\rangle \leqq 0, |y_N - z_{x,N}| \leqq C_8 hR(\sqrt{h})^2\},\\
R &= \{y \in B_N(x, \sqrt{h}R(\sqrt{h})) \cap B_N(x, \operatorname{dist}(x, \partial\Omega) + C_8\sqrt{h}R(\sqrt{h}))^c\\
&\qquad \mid \langle a, y-x\rangle \geqq 0, \langle b, y-x\rangle \leqq 0\},\\
\widetilde{Q} &= \{y \in B_N(x, \sqrt{h}R(\sqrt{h})) \cap B_N(x, \operatorname{dist}(x, \partial\Omega) + C_8\sqrt{h}R(\sqrt{h}))^c\\
&\qquad \mid \langle a, y-x\rangle \leqq 0\},\\
\widetilde{R} &= \{y \in B_N(x, \sqrt{h}R(\sqrt{h})) \cap B_N(x, \operatorname{dist}(x, \partial\Omega) + C_8\sqrt{h}R(\sqrt{h}))^c\\
&\qquad \mid \langle a, y-x\rangle \geqq 0, \langle b, y-x\rangle \geqq 0\}.
\end{aligned}$$

Then applying Lemmas 3.12 and 3.13, we have, for some $h_7 > 0$,

$$\begin{aligned}
&\int_{\Omega^c \cap B_N(x, \sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)\chi_{\{\varphi < \widetilde{\mu}\}}(y)dy \leqq \left(\int_Q + \int_{\widetilde{Q}}\right) f(y)dy\\
&\quad \leqq \left(\int_R + \int_{\widetilde{R}}\right) f(y)dy = \int_{\Omega^c \cap B_N(x, \sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)\chi_{\{\varphi \geqq \widetilde{\mu}\}}(y)dy
\end{aligned}$$

for all $h \in (0, h_7)$. Therefore, there exists a $C_9 > 0$ independent of ε, $h > 0$, and $x \in B_N(z, r_0)$ such that

$$\begin{aligned}
&\int_\Omega f_{\sqrt{h}}(y-x)\chi_{\{\varphi \geqq \widetilde{\mu} + C_9\varepsilon h\}}(y)dy + \varepsilon\sqrt{h} - \sqrt{h}\omega(R(\sqrt{h}))\\
&\quad \leqq \int_{\Omega \cap B_N(x, \sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)\chi_{\{\varphi \geqq \widetilde{\mu}\}}(y)dy\\
&\quad = \int_{B_N(x, \sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)\chi_{\{\varphi \geqq \widetilde{\mu}\}}(y)dy\\
&\qquad - \int_{\Omega^c \cap B_N(x, \sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)\chi_{\{\varphi \geqq \widetilde{\mu}\}}(y)dy\\
&\quad \leqq \int_{B_N(x, \sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)\chi_{\{\varphi < \widetilde{\mu}\}}(y)dy\\
&\qquad - \int_{\Omega^c \cap B_N(x, \sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)\chi_{\{\varphi < \widetilde{\mu}\}}(y)dy\\
&\quad = \int_{\Omega \cap B_N(x, \sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)\chi_{\{\varphi < \widetilde{\mu}\}}(y)dy\\
&\quad < \int_\Omega f_{\sqrt{h}}(y-x)\chi_{\{\varphi < \widetilde{\mu} + C_9\varepsilon h\}}(y)dy.
\end{aligned}$$

Hence taking $\varepsilon_2 > 0$ small and $0 < h_8 \leqq \min\{h_i \mid i = 5, 6, 7\}$, we get $\lambda \leqq \widetilde{\mu} + C_9\varepsilon h \leqq \widetilde{\lambda} + (C_2 + C_9)\varepsilon h$ for all $\varepsilon \in (0, \varepsilon_2)$ and $h \in (0, h_8)$.

Case 2. $\{y \in B_N(x, \sqrt{h}R(\sqrt{h})) \mid \varphi(y) = \widetilde{\mu}\} \cap \partial\Omega = \emptyset$.

Subcase 2-1. $\{y \in B_N(x, \sqrt{h}R(\sqrt{h})) \mid \varphi(y) = \widetilde{\mu}\} \subset \Omega^c$.

It follows from Lemma 3.7 that for some $C_{11} > 0$ independent of ε, $h > 0$, and $x \in B_N(z, r_0)$,

$$\operatorname{dist}(x, \partial\Omega) \leqq C_{10}h \quad \text{for all } h \in (0, h_1). \tag{3.14}$$

It is easily seen that

$$\int_\Omega f_{\sqrt{h}}(y-x)\chi_{\{\varphi\geqq\widetilde{\mu}\}}(y)dy \leqq \sqrt{h}\omega(R(\sqrt{h})),$$
$$\int_{\Omega\cap B_N(x,\sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)\chi_{\{\varphi<\widetilde{\mu}\}}(y)dy = \int_{\Omega\cap B_N(x,\sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)dy.$$

By the smoothness of $\partial\Omega$ we can show that

$$\partial\Omega\cap B_N(x,\sqrt{h}R(\sqrt{h}))\subset\{y\in B_N(x,\sqrt{h}R(\sqrt{h}))\mid |y_N-z_{x,N}|\leqq C_{11}hR(\sqrt{h})^2\},$$

where $C_{11}>0$ is independent of ε, $h>0$, and $x\in B_N(z,r_0)$, and z_x is a unique point satisfying $\operatorname{dist}(x,\partial\Omega)=|x-z_x|$. Thus it follows from (3.9), (3.14), and (A.4) that there exists an $h_9>0$ such that, for all $h\in(0,h_9)$, we observe

$$\begin{aligned}
&\int_{\Omega\cap B_N(x,\sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)dy\\
&\quad\geqq\frac{1}{2}\int_{B_N(0,\sqrt{h}R(\sqrt{h}))} f(y)dy-\int_{-C_{11}\sqrt{h}(R(\sqrt{h})+1)}^{C_{11}\sqrt{h}(R(\sqrt{h})+1)}\int_{B_{N-1}(0,R(\sqrt{h}))} f(y',y_N)dy'dy_N\\
&\quad\geqq\frac{1}{2}\int_{\mathcal{R}^N} f(y)dy-\sqrt{h}\omega(R(\sqrt{h}))-\int_{-C_{11}\sqrt{h}(R(\sqrt{h})+1)}^{C_{11}\sqrt{h}(R(\sqrt{h})+1)}\left(\int_{\mathcal{R}^{N-1}} f(y',0)dy'+1\right)dy_N\\
&\quad\geqq\sqrt{h}\omega(R(\sqrt{h})).
\end{aligned}$$

Therefore, we obtain $\lambda\leqq\widetilde{\mu}\leqq\widetilde{\lambda}+C_2\varepsilon h$ for all $\varepsilon\in(0,\varepsilon_1)$ and $h\in(0,h_1\wedge h_9)$.

Subcase 2-2. $\{y\in B_N(x,\sqrt{h}R(\sqrt{h}))\mid\varphi(y)=\widetilde{\mu}\}\subset\Omega$ and $\operatorname{dist}(x,\partial\Omega)\leqq\sqrt{h}R_1$.

As in Subcase 1-1, we observe that there exist a $C_{12}>0$ and an $h_{10}>0$ such that

$$\int_{\Omega\cap B_N(x,\sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)\chi_{\{\varphi\geqq\widetilde{\mu}\}}(y)dy\leqq\frac{1}{2}\int_{B_N(x,\sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)dy-C_{12}$$

for all $h\in(0,h_{10})$. It follows from this inequality that

$$\begin{aligned}
\int_\Omega f_{\sqrt{h}}(y-x)\chi_{\{\varphi\geqq\widetilde{\mu}\}}(y)dy&\leqq\frac{1}{2}\int_{B_N(x,\sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)dy-C_{12}+\sqrt{h}\omega(R(\sqrt{h}))\\
&\leqq\int_\Omega f_{\sqrt{h}}(y-x)\chi_{\{\varphi<\widetilde{\mu}\}}(y)dy
\end{aligned}$$

for any $h\in(0,h_{10})$. Thus $\lambda\leqq\widetilde{\mu}\leqq\widetilde{\lambda}+C_2\varepsilon h$.

Subcase 2-3. $\{y\in B_N(x,\sqrt{h}R(\sqrt{h}))\mid\varphi(y)=\widetilde{\mu}\}\subset\Omega$ and $\operatorname{dist}(x,\partial\Omega)>\sqrt{h}R_1$.

Note that

$$\begin{aligned}
&\int_{\Omega\cap B_N(x,\sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)\chi_{\{\varphi\geqq\widetilde{\mu}\}}(y)dy\\
&\quad\leqq\int_{\Omega\cap B_N(x,\sqrt{h}R(\sqrt{h}))} f_{\sqrt{h}}(y-x)\chi_{\{\varphi<\widetilde{\mu}\}}(y)dy.
\end{aligned}$$

Then computing as in Subcase 1-2, we have

$$\begin{aligned}
&\int_\Omega f_{\sqrt{h}}(y-x)\chi_{\{\varphi\geqq\widetilde{\mu}+C_{13}\varepsilon h\}}(y)dy\\
&\quad\leqq\int_\Omega f_{\sqrt{h}}(y-x)\chi_{\{\varphi<\widetilde{\mu}+C_{13}\varepsilon h\}}(y)dy+\sqrt{h}\omega(R(\sqrt{h}))-\varepsilon\sqrt{h}
\end{aligned}$$

for some $C_{13} > 0$ independent of ε, $h > 0$, and $x \in B_N(z, r_0)$. Taking $h_{11} > 0$ small, we have $\lambda \leqq \widetilde{\mu} + C_{13}\varepsilon h \leqq \widetilde{\lambda} + (C_2 + C_{13})\varepsilon h$ for all $h \in (0, h_{11})$.

Consequently, taking $C > 0$ large and ε_0, $h_0 > 0$, small enough, we complete the proof. □

4. Proof of Theorem 2.1. Before proving our main result, we prepare some lemmas.

LEMMA 4.1. *Let $\varepsilon > 0$ and $g \in C^2(\overline{\Omega})$. Assume $\partial g/\partial n > 0$ (resp., $\partial g/\partial n < 0$) on $\partial\Omega$. Then there exists a $\omega \in C(\mathcal{R}^+; \mathcal{R}^+)$, $\omega(0) = 0$, such that*

$$\sup_{x\in\overline{\Omega}, m\in\mathcal{N}} (u^m(t,x) - g(x)) \leqq \omega(t), \left(\text{resp., } \inf_{x\in\overline{\Omega}, m\in\mathcal{N}} (u^m(t,x) - g(x)) \geqq -\omega(t)\right)$$

for all $x \in \overline{\Omega}$ and $t \geqq 0$.

Proof. We may assume $\partial g/\partial n > 0$ on $\partial\Omega$ because we can prove the result similarly in the case $\partial g/\partial n < 0$ on $\partial\Omega$.

We prove that there exist a $C > 0$ and an $h_0 > 0$ such that

$$G_h g(x) \leqq g(x) + Ch \quad \text{for all } x \in \overline{\Omega},\ h \in (0, h_0). \tag{4.1}$$

Once we have this inequality, we use it iteratively to obtain our desired result with $\omega(t) = Ct$. Fix $z \in \overline{\Omega}$.

Case 1. $Dg(z) \neq 0$.

By using Lemma 3.1 with $\varepsilon = 1$, we see that there exists a $\delta_1 > 0$ such that

$$G_h g(x) \leqq g(x) + (-F(Dg(z), D^2g(z)) + 1)h$$

for all $x \in B_N(z, \delta_1) \cap \overline{\Omega}$ and $h \in (0, \delta_1)$.

Case 2. $Dg(z) = 0$.

Note that $z \in \Omega$ because we assume $\partial g/\partial n > 0$ on $\partial\Omega$. Since $g \in C^2(\overline{\Omega})$, we can find a $\delta_2 > 0$ such that

$$|g(x) - g(z)| \leqq \frac{1}{2}\|D^2g\|_\infty |x - z|^2 \quad \text{for all } x \in B_N(z, \delta_2) \subset \Omega. \tag{4.2}$$

Moreover, taking $C_1 = 2\|g\|_\infty/\delta_2^2 + \|D^2g\|_\infty/2$, we get

$$g(x) \leqq g(z) + C_1|x - z|^2 \quad \text{for all } x \in \overline{\Omega}.$$

Thus we have

$$G_h g(x) \leqq g(z) + C_1 G_h(|x - z|^2) \quad \text{for all } x \in \overline{\Omega}.$$

By a similar argument in Ishii [6] and Ishii, Pires, and Souganidis [7], we obtain

$$G_h(|x - z|^2) \leqq |x - z|^2 + C_2 h \quad \text{for all } x \in B_N(z, \delta_3),\ h \in (0, \delta_3),$$

for some $C_2 > 0$ and $\delta_3 > 0$. Hence, by (4.2) we get

$$\begin{aligned} G_h g(x) &\leqq g(z) + C_1(|x - z|^2 + C_2 h) \\ &\leqq g(x) + \left(C_1 + \frac{1}{2}\|D^2g\|_\infty\right)(|x - z|^2 + C_2 h) \\ &\qquad \text{for all } x \in B_N(z, \delta_2 \wedge \delta_3),\ h \in (0, \delta_2 \wedge \delta_3). \end{aligned}$$

Consequently, for any $z \in \overline{\Omega}$, there exist C_3, $\delta_4 > 0$ such that

$$G_h g(x) \leqq g(x) + C_3 h \quad \text{for all } x \in B_N(z, \delta_4),\ h \in (0, \delta_4).$$

Since $\overline{\Omega}$ is compact, we have (4.1). Thus the proof is completed. □

Define $\overline{u}(t,x)$ and $\underline{u}(t,x)$ by

$$\overline{u}(t,x) = \lim_{\varepsilon \to 0} \sup\{u^m(s,y) \mid m \geqq 1/\varepsilon, y \in \overline{\Omega}, 0 \leqq s < T, |s-t| + |y-x| \leqq \varepsilon\},$$
$$\underline{u}(t,x) = \lim_{\varepsilon \to 0} \inf\{u^m(s,y) \mid m \geqq 1/\varepsilon, y \in \overline{\Omega}, 0 \leqq s < T, |s-t| + |y-x| \leqq \varepsilon\}.$$

LEMMA 4.2. *Let $g \in C(\overline{\Omega})$. Then $\overline{u}(0,x) = \underline{u}(0,x) = g(x)$ for all $x \in \overline{\Omega}$.*

Proof. Take a sequence $\{g_k\} \subset C^2(\overline{\Omega})$ satisfying

$$\|g_k - g\|_{C(\overline{\Omega})} \leqq \frac{1}{k}, \quad \frac{\partial g_k}{\partial n} > 0 \quad \text{on} \quad \partial\Omega.$$

Define $u^{m,k}$ as (2.5) with g replacing g_k. Then Lemma 4.1 implies that, for each $n \in \mathcal{N}$, there exists an $\omega_k \in C(\mathcal{R}^+; \mathcal{R}^+)$, $\omega_k(0) = 0$, such that

$$u^{m,k}(t,y) - g_n(y) \leqq \omega_k(t) \quad \text{for all } y \in \overline{\Omega},\ t \geqq 0.$$

Since G_h is nonexpansive and $\|g_k - g\|_{C(\overline{\Omega})} \leqq 1/k$, we observe that

$$|u^m(t,y) - u^{m,k}(t,y)| \leqq \frac{1}{k} \quad \text{for all } y \in \overline{\Omega},\ t \geqq 0,\ k \in \mathcal{N}.$$

Hence we get

$$u^m(t,y) - g(y) \leqq \omega_k(t) + \frac{2}{k}.$$

Fix $x \in \overline{\Omega}$. Letting $m \to +\infty$, $y \to x$, $t \to 0$, and then $k \to +\infty$, we have

$$\overline{u}(0,x) \leqq g(x).$$

Since we can show $\underline{u}(0,x) \geqq g(x)$ by a similar argument and $\underline{u}(0,x) \leqq \overline{u}(0,x)$, we have the result. □

Proof of Theorem 2.1. By using Lemma 3.1, we can show that, for any $\varphi \in C^2(\overline{\Omega})$, if $x \in \Omega$ and $D\varphi(x) \neq 0$ or $x \in \partial\Omega$ and $(\partial\varphi/\partial n)(x) > 0$, then

$$\limsup_{h \to 0}{}^{*} \frac{G_h\varphi(x) - \varphi(x)}{h} \leqq -F_*(D\varphi(x), D^2\varphi(x)), \tag{4.3}$$

and if $x \in \Omega$ and $D\varphi(x) \neq 0$ or $x \in \partial\Omega$ and $(\partial\varphi/\partial n)(x) < 0$, then

$$\lim_{h \to 0}{}_{*} \frac{G_h\varphi(x) - \varphi(x)}{h} \geqq -F^*(D\varphi(x), D^2\varphi(x)),$$

where

$$\lim_{h \to 0}{}^{*} \psi_h(x) = \lim_{r \to 0} \sup\{\psi_h(y) \mid 0 < h < r, |y-x| < r\},$$
$$\lim_{h \to 0}{}_{*} \psi_h(x) = -\lim_{h \to 0}{}^{*}(-\psi_h(x)).$$

We prove $\overline{u}$ is a viscosity subsolution of (2.1). Fix $\varphi \in C^2((0,T) \times \overline{\Omega})$ and assume $\overline{u} - \varphi$ takes its maximum at (t_0, x_0). If $x_0 \in \partial\Omega$ and $(\partial\varphi/\partial n)(t_0, x_0) \leqq 0$, then we have nothing to prove. Thus we assume $x_0 \in \Omega$ or $x_0 \in \partial\Omega$ and $(\partial\varphi/\partial n)(t_0, x_0) > 0$. Then, by (4.3) and the same argument as in Ishii [6] or Ishii, Pires, and Souganidis [7], we conclude that $\overline{u}$ is a viscosity subsolution of (2.1). It can be proved similarly that $\underline{u}$ is a viscosity supersolution of (2.1). Therefore, Lemma 4.2 and the comparison principle due to Giga and Sato [5] yield $\overline{u} = \underline{u}$ in $[0,T) \times \overline{\Omega}$. Hence, by the results in Crandall, Ishii, and Lions [3, section 6], the proof is completed. □

REFERENCES

[1] G. Barles and C. Georgelin, *A simple proof of convergence for an approximation scheme for computing motion by mean curvature*, SIAM J. Numer. Anal., 32 (1995), pp. 484–500.

[2] J. Bence, B. Merriman, and S. Osher, *Diffusion generated motion by mean curvature*, in Computational Crystal Growers Workshop, J. E. Taylor, ed., Sel. Lectures Math., AMS, Providence, RI, 1992.

[3] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.

[4] L. C. Evans, *Convergence of an algorithm for mean curvature motion*, Indiana Univ. Math. J., 42 (1993), pp. 533–557.

[5] Y. Giga and M.-H. Sato, *Neumann problem for singular degenerate parabolic equations*, Differential Integral Equations, 6 (1993), pp. 1217–1230.

[6] H. Ishii, *A generalization of the Bence, Merryman and Osher algorithm for motion by mean curvature*, in Curvature Flows and Related Topics, A. Damlamian, J. Spruck, and A. Visintin, eds., Gakuto Internat. Ser. Math. Sci. Appl. 5, Gakkōtosho, Tokyo, 1995, pp. 111–127.

[7] H. Ishii, G. E. Pires, and P. E. Souganidis, *Threshold dynamics type approximiation schemes for propagating fronts*, J. Math. Soc. Japan, 50 (1999), pp. 267–308.

[8] M. H. Sato, *Interface evolution with Neumann boundary condition*, Adv. Math. Sci. Appl., 4 (1994), pp. 249–264.

SIAM J. Math. Anal.
Vol. 33, No. 2, pp. 390–410

ZERO VISCOSITY LIMIT OF THE OSEEN EQUATIONS IN A CHANNEL*

MARIA CARMELA LOMBARDO[†] AND MARCO SAMMARTINO[†]

Abstract. Oseen equations in the channel are considered. We give an explicit solution formula in terms of the inverse heat operators and of projection operators. This solution formula is used for the analysis of the behavior of the Oseen equations in the zero viscosity limit. We prove that the solution of Oseen equations converges in $W^{1,2}$ to the solution of the linearized Euler equations outside the boundary layer and to the solution of the linearized Prandtl equations inside the boundary layer.

Key words. Oseen equations, solution formula, zero viscosity limit, boundary layer, asymptotic expansion

AMS subject classifications. 35Q30, 76D07, 76D10, 35C20

PII. S0036141000372015

1. Introduction.

In this paper we shall be concerned with the zero viscosity limit of the time dependent incompressible Oseen type equations in a two-dimensional channel.

Oseen equations are a simplified mathematical model in treating incompressible viscous fluids and their study should shed some light on the much more complicated physical phenomena described by the nonlinear Navier–Stokes equations.

The system we shall be concerned with is the following:

$$(\partial_t - \nu\Delta + U\partial_x)\boldsymbol{u} + \boldsymbol{\nabla} p = \boldsymbol{f}, \tag{1.1}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{u} = 0, \tag{1.2}$$

$$\gamma^-\boldsymbol{u} = \mathbf{g}^-, \tag{1.3}$$

$$\gamma^+\boldsymbol{u} = \mathbf{g}^+, \tag{1.4}$$

$$\boldsymbol{u}(t=0) = \boldsymbol{u}_0. \tag{1.5}$$

In the above equations $\boldsymbol{u} = (u(x,y,t), v(x,y,t))$ is the unknown vector field, p is the unknown pressure, $x \in \mathbb{R}$ denotes the tangential variable, $-1 \le y \le 1$ is the normal variable, $t \ge 0$ is the time, γ is the trace operator, i.e., $\gamma^-\boldsymbol{u} = \boldsymbol{u}(x,-1,t)$ and $\gamma^+\boldsymbol{u} = \boldsymbol{u}(x,1,t)$, ν is the viscosity coefficient, and $\mathbf{g}^-, \mathbf{g}^+$, and $\boldsymbol{u}_0$ are the boundary and the initial data, respectively.

Equation (1.1) is obtained by linearizing the Navier–Stokes equation around the velocity flow $(U,0)$, where U is assumed to be constant. Equation (1.2) is the incompressibility condition. Equations (1.3) and (1.4) are the boundary conditions, while (1.5) is the initial condition.

The analysis of the zero viscosity limit of (1.1)–(1.5) (with homogeneous boundary data $\mathbf{g}^+ = \mathbf{g}^- = 0$) was performed by Temam and Wang in [4], [5], and [6]. In [6] the authors considered the more general case of the three-dimensional Navier–Stokes equations linearized around the flow $(U_1(z), U_2(z), 0)$. Without explicitly solving the

*Received by the editors May 3, 2000; accepted for publication (in revised form) February 22, 2001; published electronically July 19, 2001.
http://www.siam.org/journals/sima/33-2/37201.html

[†]Dipartimento di Matematica, University of Palermo, Via Archirafi 34, 90123 Palermo, Italy (lombardo@dipmat.math.unipa.it, marco@dipmat.math.unipa.it). The research of the first author was partially supported by the INDAM. The research of the second author was partially supported by the GNFM-INDAM.

equations and under the assumption of sufficient smoothness of the initial data, they proved that the solution can be decomposed in the following form:

$$\boldsymbol{u} = \boldsymbol{u}^0 + \tilde{\boldsymbol{\theta}}^{\boldsymbol{\varepsilon}} + \nu^{1/4}\boldsymbol{w} \qquad \text{with} \qquad \boldsymbol{w} \in L^2\left([0,T];\ \left(W^{1,2}(\Omega)\right)^3\right),$$

where $\boldsymbol{u}^0$ is the inviscid solution and $\tilde{\boldsymbol{\theta}}^{\boldsymbol{\varepsilon}}$ is a corrector which takes into account the mismatch of the boundary conditions between the inviscid part and the overall solution.

In this paper we shall write the solution of (1.1)–(1.5) as the sum of an inviscid part, two Prandtl parts, and an error term, i.e.,

$$\boldsymbol{u} = \boldsymbol{u}^E + \boldsymbol{u}^{P(-)} + \boldsymbol{u}^{P(+)} + \nu^{1/2}\boldsymbol{w}\ . \tag{1.6}$$

In the above decomposition $\boldsymbol{u}^E$ is the inviscid (Euler) part describing the flow away from boundaries; $\boldsymbol{u}^{P(-)}$ and $\boldsymbol{u}^{P(+)}$ are the Prandtl parts describing the flow close to the two boundaries and decaying exponentially away from the two boundaries. Our main result is the following theorem.

Theorem 1.1 (informal statement). *Let $\boldsymbol{u}$ be the solution of* (1.1)–(1.6)*. If the initial data $\boldsymbol{u}_0$ is in H^4, and if the boundary data $\boldsymbol{g}^{\pm}$ are sufficiently regular and are such that the normal components are integrable (in the sense specified by Definition* 2.3 *below), then*

$$\nu^{1/4}\boldsymbol{w} \in L^\infty\left([0,T]\,; W^{1,2}\right)$$

with norm independent of ν.

Therefore the main advantage with respect to the results of [6] is that here we prove that the norm of the correction term is $O(\nu^{1/4})$ in $L^\infty\left([0,T]; W^{1,2}\right)$, while there an analogous estimate is proved with a nondivergence-free boundary layer corrector. We also allow nonzero boundary data.

We shall achieve this result through the explicit solution formula for the Oseen equation. This explicit representation formula is, we believe, of independent interest.

The paper is organized as follows: in section 2 we shall introduce the functional setting. In section 3 the convection-diffusion operators in the channel are defined and we shall give some estimate in the appropriate function spaces. In section 4 we shall solve the Oseen equations with boundary data, zero initial data, and zero source term. The solution will be written in the form of an infinite series: the norm of the generic term of the series will be shown to be exponentially decaying. Convergence of the series will follow and the norm of the solution will be proved to be bounded in terms of the norm of the boundary data. In section 5 the complete Oseen system is solved introducing the projection operator onto the divergence-free function space and estimates of the solution in terms of the data are given. In section 6 we shall finally analyze the vanishing viscosity limit of the Oseen equations. The main results of this paper are formally stated in Theorem 6.5 and Corollary 6.1.

2. The function spaces.

In this section we define some function spaces we shall be using throughout the paper. We introduce the notation $\boldsymbol{\Omega} \equiv \mathbb{R} \times [-1,1]$. Here and in the rest of the paper $l \geq 2$.

Definition 2.1. *$H^l(\mathbb{R})$ is the set of all functions $f(x)$ such that*

(i) $\frac{d^j}{dx^j} f \in L^2(\mathbb{R})$*, where $j \leq l$.*

We shall denote the usual norm in $H^l(\mathbb{R})$ with $|f|_l$.

Definition 2.2. *H'^{l}_T is the set of all functions $f(x,t)$ such that*

(i) *$\partial_t^j f(x,t) \in L^\infty([0,T], H^{l-j}(\mathbb{R}))$ and $j \le l$.*

The norm of $f \in H'^{l}_T$ is given by

$$|f|_{l,T} = \sum_{j_1+j_2 \le l} \sup_{0 \le t \le T} \|\partial_t^{j_1} \partial_x^{j_2} f(\cdot,t)\|_{L^2(\mathbb{R})} \ .$$

Definition 2.3. *$\boldsymbol{H}'^{l}_T$ is the set of all functions $\boldsymbol{f} = (f_\tau, f_N)$ such that*

(i) *$f_\tau \in H'^{l}_T$;*

(ii) *$|\xi|^{-1} f_N \in H'^{l+1}_T$.*

The norm of $\boldsymbol{f} \in \boldsymbol{H}'^{l}_T$ is given by

$$\boldsymbol{|}\, \boldsymbol{f} \,\boldsymbol{|}_{\,l,T} = |f_\tau|_{l,T} + ||\xi|^{-1} f_N|_{l,T} \ .$$

In the above definition ξ is the dual Fourier variable of x, and $|\xi|^{-1}$ has to be understood as a pseudodifferential operator. The space $\boldsymbol{H}'^{l}_T$ is the space to which all boundary data we shall deal with in the rest of the paper will belong. The hypothesis on the normal component is an *integrability* hypothesis.

Definition 2.4. *H^l is the set of all functions $f(x,y)$ such that*

(i) *$\partial_x^i \partial_y^j f(x,y) \in L^2(\boldsymbol{\Omega})$ for $i+j \le l$.*

The norm of f is given by

$$|f|_l = \sum_{i+j \le l} \|\partial_x^i \partial_y^j f(\cdot,\cdot)\|_{L^2(\boldsymbol{\Omega})} \ .$$

Definition 2.5.

H^l_T is the set of all functions $f(x,y,t)$ such that

- *$\partial_t^j f(x,y,t) \in L^\infty([0,T], H^{l-j})$ for $j \le l$.*

The norm of $f \in H^l_T$ is given by

$$|f|_{l,T} = \sum_{j_1+j_2+j_3 \le l} \sup_{0 \le t \le T} \|\partial_t^{j_1} \partial_x^{j_2} \partial_y^{j_3} f(\cdot,\cdot,t)\|_{L^2(\boldsymbol{\Omega})}.$$

All the above spaces are the natural ambient spaces for the Euler equations. We now introduce the ambient spaces for Prandtl equations. All the functions belonging to these spaces depend on the normal scaled variable $Y = y/\varepsilon$. We require differentiability with respect to this variable only up to the second order. We first introduce the spaces $K^{l,\mu(+)}$ and $K^{l,\mu(-)}$. The functions in these spaces are defined in the half plane $Y \le 1/\varepsilon$ and $Y \ge -1/\varepsilon$, respectively, and decay exponentially fast away from $Y = 1/\varepsilon$ and $Y = -1/\varepsilon$, respectively. In what follows $\mu > 0$.

Definition 2.6. *$K^{l,\mu(\pm)}$ is the set of all functions $f^\pm(x,Y)$ defined for $-\infty < \pm Y \le 1/\varepsilon$, and such that*

(i) *$\partial_x^{j_1} \partial_Y^{j_2} f^\pm(x,Y) \in L^2(\mathbb{R})$, with $j_2 \le 2$ and $j_1 + j_2 \le l$,*

(ii) *$\sup_{-\infty < \pm Y \le \frac{1}{\varepsilon}} e^{\mu(\frac{1}{\varepsilon} \mp Y)} \|\partial_x^{j_1} \partial_Y^{j_2} f^\pm(\cdot,Y)\|_{L^2} < \infty$, where $j_2 \le 2$ and $j_1 + j_2 \le l$.*

The norm is given by

$$|f|_{l,\mu(\pm)} = \sum_{j_2 \le 2} \sum_{j_1+j_2 \le l} \sup_{-\infty < \pm Y \le \frac{1}{\varepsilon}} e^{\mu(\frac{1}{\varepsilon} \mp Y)} \|\partial_x^{j_1} \partial_Y^{j_2} f^\pm(\cdot,Y)\|_{L^2} \ .$$

Definition 2.7. *$K^{l,\mu}$ is the set of functions $f(x,Y)$, defined for $-1/\varepsilon \le Y \le 1/\varepsilon$, such that*

$$f = f^+ + f^-,$$

where f^+ and f^- are restrictions to $-1/\varepsilon \le Y \le 1/\varepsilon$ of functions in $K^{l,\mu(+)}$ and $K^{l,\mu(-)}$, respectively. The norm is given by

$$|f|_{l,\mu}=\sum_{j_2\le 2}\sum_{j_1+j_2\le l}\left[\sup_{0<Y\le\frac{1}{\varepsilon}} e^{\mu(\frac{1}{\varepsilon}-Y)}\|\partial_x^{j_1}\partial_Y^{j_2}f(\cdot,Y)\|_{L^2}+\sup_{-\frac{1}{\varepsilon}<Y\le 0} e^{\mu(\frac{1}{\varepsilon}+Y)}\|\partial_x^{j_1}\partial_Y^{j_2}f(\cdot,Y)\|_{L^2}\right].$$

We now introduce the dependence on time. We require differentiability with respect to time only up to the first order: one time derivative is equivalent to two space derivatives.

Definition 2.8.
$K_T^{l,\mu(\pm)}$ is the set of all functions $f^\pm(x,Y,t)$ such that
(i) *$f \in L^\infty([0,T],K^{l,\mu(\pm)})$,*
(ii) *$\partial_t\partial_x^j f \in L^\infty([0,T],K^{0,\mu(\pm)})$ with $j \le l-2$.*
The norm is given by

$$\begin{aligned}|f|_{l,\mu,T(\pm)} = &\sum_{0\le j_2\le 2}\sum_{j_1\le l-2}\sup_{0\le t\le T}\sup_{-\infty<\pm Y\le\frac{1}{\varepsilon}} e^{\mu(\frac{1}{\varepsilon}\mp Y)}\|\partial_x^{j_1}\partial_Y^{j_2}f^\pm(\cdot,Y,t)\|_{L^2}\\ &+\sum_{j\le l-2}\sup_{0\le t\le T}\sup_{-\infty<\pm Y\le\frac{1}{\varepsilon}} e^{\mu(\frac{1}{\varepsilon}\mp Y)}\|\partial_t\partial_x^j f^\pm(\cdot,Y,t)\|_{L^2}\ .\end{aligned}$$

Definition 2.9. *$K_T^{l,\mu}$ is the set of functions $f(x,Y,t)$, defined for $-1/\varepsilon \le Y \le 1/\varepsilon$, such that*

$$f = f^+ + f^-\ ,$$

where f^+ and f^- are restrictions to $-1/\varepsilon \le Y \le 1/\varepsilon$ of functions in $K_T^{l,\mu(+)}$ and $K_T^{l,\mu(-)}$, respectively. The norm is given by

$$\begin{aligned}|f|_{l,\mu,T} = &\sum_{0\le j_2\le 2}\sum_{j_1\le l-2}\sup_{0\le t\le T}\sup_{0<Y\le\frac{1}{\varepsilon}} e^{\mu(\frac{1}{\varepsilon}-Y)}\|\partial_x^{j_1}\partial_Y^{j_2}f(\cdot,Y,t)\|_{L^2}\\ &+\sum_{j\le l-2}\sup_{0\le t\le T}\sup_{0<Y\le\frac{1}{\varepsilon}} e^{\mu(\frac{1}{\varepsilon}-Y)}\|\partial_t\partial_x^j f(\cdot,Y,t)\|_{L^2}\\ &+\sum_{0\le j_2\le 2}\sum_{j_1\le l-2}\sup_{0\le t\le T}\sup_{-\frac{1}{\varepsilon}<Y\le 0} e^{\mu(\frac{1}{\varepsilon}+Y)}\|\partial_x^{j_1}\partial_Y^{j_2}f(\cdot,Y,t)\|_{L^2}\\ &+\sum_{j\le l-2}\sup_{0\le t\le T}\sup_{-\frac{1}{\varepsilon}<Y\le 0} e^{\mu(\frac{1}{\varepsilon}+Y)}\|\partial_t\partial_x^j f(\cdot,Y,t)\|_{L^2}\ .\end{aligned}$$

We now introduce the ambient spaces for the error equation. All functions belonging to the following spaces are functions L^2 with respect to both tangential and normal variables. Notice that, due to the presence in the error equation of the rapidly varying terms arising from the Prandtl solution, the solution of the error equation will have a fast dependence on y. Therefore, in the following spaces, all the derivatives of order j with respect to y are weighted with $\varepsilon^j \equiv \nu^{j/2}$.

Definition 2.10. *L^l is the set of all functions $f(x,y)$ such that*

(i) *$\partial_x^{j_1}\varepsilon^{j_2}\partial_y^{j_2}f \in L^2(\mathbf{\Omega})$ with $j_2 \leq 2$ and $j_1 + j_2 \leq l$.*

The norm of $f \in L^l$ is given by

$$\|f\|_l = \sum_{j_1 \leq l} \|\partial_x^{j_1} f\|_{L^2(\mathbf{\Omega})} + \sum_{0\leq j_2 \leq 2}\sum_{j_1 \leq l-2} \|\partial_x^{j_1}\varepsilon^{j_2}\partial_y^{j_2} f\|_{L^2(\mathbf{\Omega})} \ .$$

Definition 2.11. *$L_T'^l$ is the set of all functions $f(x,t)$ such that*

(i) *$\partial_t\partial_x^j f \in L^\infty([0,T], H^l(\mathbb{R}))$ for $j \leq l-2$.*

The norm of $f \in L_T'^l$ is given by

$$\|f\|_{l,T} = \sum_{j \leq l-2} \sup_{0\leq t\leq T} \|\partial_t\partial_x^j f(\cdot,t)\|_{L^2(\mathbf{\Omega})} \ .$$

Definition 2.12. *L_T^l is the set of all functions f(x,y,t) such that*

(i) *$f \in L^\infty([0,T], L^l)$,*

(ii) *$\partial_t\partial_x^j f \in L^\infty([0,T], L^0)$ with $j \leq l-2$.*

The norm of $f \in L_T^l$ is given by

$$\|f\|_{l,T} = \sum_{0\leq j_2\leq 2}\sum_{j_1\leq l-2} \sup_{0\leq t\leq T} \|\partial_x^{j_1}\varepsilon^{j_2}\partial_y^{j_2} f(\cdot,\cdot,t)\|_{L^2(\mathbf{\Omega})} + \sum_{j\leq l-2}\sup_{0\leq t\leq T}\|\partial_t\partial_x^j f(\cdot,\cdot,t)\|_{L^2(\mathbf{\Omega})} \ .$$

3. The convection-diffusion operators in the channel. In this section we shall give an explicit representation of the convection-diffusion operators F_0, F_1, and F_2. These operators solve the convection-diffusion equation in the channel with initial data, boundary data, and source term, respectively. Our explicit representation is given in terms of the action on the Fourier transform in the tangential variable. These operators will be used in the construction of the explicit solution of the Oseen equations. In what follows ξ will always denote the dual Fourier variable of x:

$$\hat{f}(\xi') = \frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{+\infty} dx\, f(x) e^{-i\xi' x} \ .$$

In the rest of the paper we shall not distinguish between a function and its Fourier transform.

The convection-diffusion equation in the channel writes as follows:

$$(\partial_t - \nu\Delta + U\partial_x)u = f \ , \tag{3.1}$$

$$\gamma^+ u = g^+ \ , \tag{3.2}$$

$$\gamma^- u = g^- \ , \tag{3.3}$$

$$u|_{t=0} = u_0 \ . \tag{3.4}$$

The operator F_0 solves the above equations with initial data and zero boundary data and zero source term. If we now introduce the θ-function (see [8, Chapter V]),

$$\theta(y,t) = \sum_{n=-\infty}^{\infty} k(y+4n,t), \qquad -\infty < y < \infty \ ,$$

where

$$k(y,t) = \frac{e^{-y^2/4\nu t}}{(4\pi\nu t)^{1/2}} \ ,$$

one can easily see that the explicit expression of F_0 is given by

$$F_0 u_0 \quad = e^{-[\nu\xi^2 + i\xi U]t} \int_{-1}^{+1} dy' \, [\theta(y-y',t) - \theta(y+y'+2,t)] \, u_0(\xi,y') \, . \tag{3.5}$$

The operator F_1 solves the convection-diffusion equations in the channel with boundary data, zero initial data, and zero source term: If we introduce the φ-function (see [8])

$$\varphi(y,t) = \sum_{n=-\infty}^{+\infty} h(y+4n,t), \qquad -\infty < y < \infty \, ,$$

where

$$h(y,t) = \frac{y}{t} \frac{e^{-y^2/(4\nu t)}}{\sqrt{4\pi t\nu}} \, ,$$

one can easily verify that the operator F_1 has the following expression:

$$\begin{aligned} F_1(g^+,g^-) = & \int_0^t ds \, e^{-(\nu\xi'^2 + i\xi' U)(t-s)} \varphi(y+1,t-s) g^-(\xi',s) \\ & + \int_0^t ds \, e^{-(\nu\xi'^2 + i\xi' U)(t-s)} \varphi(1-y,t-s) g^+(\xi',s) \, . \end{aligned} \tag{3.6}$$

Finally, the operator F_2 solves the convection-diffusion equations with source term, zero initial data, and zero boundary data. One can give the following explicit expression for F_2:

$$F_2 f = \int_0^t ds e^{-[\nu\xi^2 + i\xi U](t-s)} \int_{-1}^{+1} dy' \, [\theta(y-y',t-s) - \theta(y+y'+2,t-s)] \, f(\xi,y',s) \, . \tag{3.7}$$

We now give some estimates on the above operators. In the propositions below c is a constant that does not depend on ν. The proofs of the proposition below are standard and can be achieved using the explicit representation formulas (3.5), (3.6), and (3.7) and the same arguments of [3], where the heat operators in the half plane are considered.

Proposition 3.1. *Let $u_0 \in L^l$ with $\gamma^\pm u_0 = 0$. Then $F_0 u_0 \in L^l_T$ and the following estimate holds:*

$$\|F_0 u_0\|_{l,T} \le c \|u_0\|_l \, .$$

Proposition 3.2. *Let $g^+ \in L'^l_T$, $g^- \in L'^l_T$ satisfy the compatibility condition $g^+(t=0) = g^-(t=0) = 0$. Then $F_1(g^+,g^-) \in L^l_T$ and the following estimate holds:*

$$\|F_1(g^+,g^-)\|_{l,T} \le c[\|g^+\|_{l,T} + \|g^-\|_{l,T}] \, .$$

Proposition 3.3. *Let $g^+ \in L'^l_T$, $g^- \in L'^l_T$, and $g^+(t=0) = g^-(t=0) = 0$; then $F_1(g^+,g^-) \in K^{l,\mu}_T$ and the following estimate holds:*

$$|F_1(g^+,g^-)|_{l,\mu,T} \le c[\|g^+\|_{l,T} + \|g^-\|_{l,T}] \, .$$

Proposition 3.4. *Let $f \in L^l_T$; then $F_2 f \in L^l_T$ and the following estimate holds:*

$$\|F_2 f\|_{l,T} \le c \|f\|_{l,T}.$$

The following estimate on the operator F_2 says that, if the source term is of the same order of the square root of ν, then one can use the heat kernel to increase the regularity of the solution.

PROPOSITION 3.5. *Let $f = \sqrt{\nu}h$ with $h \in L_T^{l-1}$; then $F_2 f \in L_T^l$ and the following estimate holds:*

$$\|F_2 f\|_{l,T} \leq c\|h\|_{l-1,T}.$$

4. The Oseen equations with boundary data: The explicit solution.

In this section we shall derive the explicit solution of the Oseen equations in the channel with boundary data, zero source term, and zero initial data, namely,

$$(\partial_t - \nu\Delta + U\partial_x)\boldsymbol{u} + \boldsymbol{\nabla} p = 0, \tag{4.1}$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{u} = 0, \tag{4.2}$$

$$\gamma^- \boldsymbol{u} = \mathbf{g}^-, \tag{4.3}$$

$$\gamma^+ \boldsymbol{u} = \mathbf{g}^+, \tag{4.4}$$

$$\boldsymbol{u}|_{t=0} = 0\ , \tag{4.5}$$

where $\mathbf{g}^+ = (g_\tau^+, g_N^+)$ and $\mathbf{g}^- = (g_\tau^-, g_N^-)$.

We shall express the solution of the above equations as a series. The nth (n even) term of the series solves Oseen equations in the channel with the *right* (in the sense that it cancels the boundary datum created at $y = -1$ by the $(n-1)$th term) boundary condition at $y = -1$. On the other hand, the nth term of the series creates a boundary datum at $y = +1$, which will be canceled by the $(n+1)$th term, which therefore has the *right* boundary condition at $y = +1$. Of course the main concern will be to show that this series is convergent. In fact we shall prove that, for a time T_α, this series is geometric with parameter $\alpha < 1$. Our representation for the solution of (4.1)–(4.5) will therefore be valid only up to time T_α. Notice, however, that $T_\alpha \to \infty$ when $\nu \to 0$.

The plan of the section is the following. First we introduce some pseudodifferential operators: (1) the operators $U^\pm$ (that we call the Ukai operators) that are projection operators, (2) the operator $\mathcal{L}$ solving Dirichlet problem in the strip, and (3) the projection-convection-diffusion operators $\mathcal{M}^+$ and $\mathcal{M}^-$. All these operators are the ingredients to construct, in section 4.4, the operators $\mathcal{O}^-$ and $\mathcal{O}^+$. These are operators that solve Oseen equations with the right boundary condition only at $y = -1$ and at $y = +1$, respectively. Finally, through these operators, in section 4.5 we construct the explicit solution formula.

4.1. The Ukai operators.

Here we define the following operators which are a modification of the operator U introduced in [7]:

$$U^- f = \int_{-1}^{y} dy' e^{-|\xi|(y-y')} f(y', \xi), \tag{4.6}$$

$$U^+ f = -\int_{y}^{1} dy' e^{|\xi|(y-y')} f(y', \xi). \tag{4.7}$$

These operators solve the problems

$$(\partial_y + |\xi|)U^- f = f, \tag{4.8}$$

$$\gamma^- U^- f = 0\ , \tag{4.9}$$

and

$$(\partial_y - |\xi|)U^+ f = f, \tag{4.10}$$

$$\gamma^+ U^+ f = 0\,. \tag{4.11}$$

The following proposition holds.

Proposition 4.1. *Let f be of the form $f = |\xi'|\tilde{f}$ with $\tilde{f} \in H_T^{l+1}$. Then $U^+ f \in H_T^l$, $U^- f \in H_T^l$, and the following estimates hold:*

$$\|U^+ f\|_{l,T} \leq c\|\tilde{f}\|_{l+1,T},$$
$$\|U^- f\|_{l,T} \leq c\|\tilde{f}\|_{l+1,T}.$$

The proof of the above proposition is a straightforward modification of the proof of a similar statement given in [1].

The following lemmas are crucial for our analysis.

Lemma 4.2. *Suppose $g_\tau^+ \in H_T^{\prime l}$ with $g_\tau^+|_{t=0} = 0$. Then $\gamma^- U^+ F_1(g_\tau^+, 0) \in H_T^{\prime l}$. Moreover, there exists $\alpha < 1$ and $T_\alpha > 0$, independent of g_τ^+, such that*

$$\|\gamma^- U^+ F_1(g_\tau^+, 0)\|_{l,T_\alpha} \leq \alpha\|g_\tau^+\|_{l,T}.$$

Lemma 4.3. *Suppose $g_\tau^- \in H_T^{\prime l}$ with $g_\tau^-|_{t=0} = 0$. Then $\gamma^- U^- F_1(0\,, g_\tau^-) \in H_T^{\prime l}$. Moreover, there exists $\alpha < 1$ and $T_\alpha > 0$, independent of g_τ^-, such that*

$$\|\gamma^+ U^- F_1(0, g_\tau^-)\|_{l,T_\alpha} \leq \alpha\|g_\tau^-\|_{l,T}.$$

The meaning of the estimate on the operator $U^+ F_1$ given in Lemma 4.2 is the following. Give to $U^+ F_1$ a boundary datum which is different from zero only at $y = +1$, and evaluate the norm of the trace at $y = -1$. Then, up to the (sufficiently small) time T_α, one has that this norm is strictly less than the norm of the boundary datum at $y = +1$. Lemma 4.3 admits a similar interpretation. It is natural to expect that the time T_α up to which the estimates are valid grows to infinity when the diffusivity goes to zero. This will show up clearly during the proof of the above lemmas, which are given in the appendix.

4.2. The inverse elliptic operators. We introduce the following operator $\mathcal{L} = (\mathcal{L}_\tau, \mathcal{L}_N)$:

$$\mathcal{L}_N(g_N^+, g_N^-) = \frac{\sinh[|\xi|(y+1)]}{\sinh[2|\xi|]} g_N^+ + \frac{\sinh[|\xi|(1-y)]}{\sinh[2|\xi|]} g_N^-, \tag{4.12}$$

$$\mathcal{L}_\tau(g_N^+, g_N^-) = \frac{\cosh[|\xi|(y+1)]}{\sinh[2|\xi|]} N' g_N^+ - \frac{\cosh[|\xi|(1-y)]}{\sinh[2|\xi|]} N' g_N^-, \tag{4.13}$$

where $N' = i\xi/|\xi|$. This operator has the property of being divergence-free and harmonic; moreover, one can easily see that it is the gradient of a scalar function. Therefore, it solves Oseen equations. Finally, it has the property that its normal component has g_N^+ and g_N^- as boundary conditions at $y = +1$ and $y = -1$, respectively.

We now give some estimates on this operator. In these estimates we shall always suppose that the boundary data $g_N^\pm$ are such that $|\xi|^{-1} g_N \in H_T^{\prime l+1}$. This hypothesis, which we shall refer to as *integrability* of the normal influx (in fact it is related to assuming that the total influx from the boundary is bounded), is necessary to handle the fact that the operator $\mathcal{L}$ is singular when $\xi \to 0$.

The proposition below gives an estimate of $\mathcal{L}$ in the spaces L^l_T and H^l_T. The estimate in the space L^l_T will be used in the next subsections in the construction of the Oseen operator. The estimate in the space H^l_T will be used in section 6 in the analysis of the Euler equations.

Proposition 4.4. *Let $g^\pm_N$ be such that $|\xi|^{-1}g^\pm_N \in H'^{l+1}_T$. Then $\mathcal{L}(g^+_N, g^-_N) \in L^l_T \bigcap H^l_T$. The following estimates hold:*

$$\begin{aligned}\|\mathcal{L}(g^+_N, g^-_N)\|_{l,T} &\le c\left[\||\xi|^{-1}g^+_N\|_{l+1,T} + \||\xi|^{-1}g^-_N\|_{l+1,T}\right],\\ |\mathcal{L}(g^+_N, g^-_N)|_{l,T} &\le c\left[\||\xi|^{-1}g^+_N\|_{l+1,T} + \||\xi|^{-1}g^-_N\|_{l+1,T}\right].\end{aligned}$$

The following lemmas will be useful in the estimates of the Oseen operator.

Lemma 4.5. *Let g^+_N be such that $|\xi|^{-1}g^+_N \in H'^{l+1}_T$. Then $\gamma^-\mathcal{L}_\tau(g^+_N, 0) \in L^l_T$, and the following estimate holds:*

$$\|\gamma^-\mathcal{L}_\tau(g^+_N, 0)\|_{l,T} \le \frac{1}{2}\||\xi|^{-1}g^+_N\|_{l+1,T}.$$

Lemma 4.6. *Let g^-_N be such that $|\xi|^{-1}g^-_N \in H'^{l+1}_T$. Then $\gamma^+\mathcal{L}_\tau(0, g^-_N) \in L^l_T$, and the following estimate holds:*

$$\|\gamma^+\mathcal{L}_\tau(0, g^-_N)\|_{l,T} \le \frac{1}{2}\||\xi|^{-1}g^-_N\|_{l+1,T}.$$

The meaning of Lemma 4.5 is the following. Give to the operator $\mathcal{L}_\tau$ a boundary datum which is nonzero only at $y = +1$. Then its trace at $y = -1$ is strictly less (half) than the datum at $y = +1$. A similar interpretation can be given to Lemma 4.6. The proof of the above two lemmas is a straightforward consequence of the fact that $\left|\frac{|\xi'|}{\sinh 2|\xi'|}\right| \le \frac{1}{2}$.

4.3. The projection-convection-diffusion operator. Through the operators we have introduced in the previous sections, we can define the operator $\mathcal{M}^-$. The tangential and normal component are defined as follows:

$$\begin{aligned}\mathcal{M}^-_\tau(g^-_\tau, g^-_N) &= U^-|\xi|F_1\left(0, g^-_\tau - \gamma^-\mathcal{L}_\tau(0, g^-_N)\right) + F_1\left(0, g^-_\tau - \gamma^-\mathcal{L}_\tau(0, g^-_N)\right),\\ \mathcal{M}^-_N(g^-_\tau, g^-_N) &= U^-|\xi|N'F_1\left(0, g^-_\tau - \gamma^-\mathcal{L}_\tau(0, g^-_N)\right).\end{aligned}$$

One can verify that the operator $\mathcal{M}^-$ is divergence-free, satisfies the Oseen equations, and has zero normal boundary condition at $y = -1$:

$$\begin{aligned}(\partial_t - \nu\Delta + U\partial_x)\,\mathcal{M}^-(g^-_\tau, g^-_N) + \boldsymbol{\nabla} p &= 0,\\ \boldsymbol{\nabla}\cdot\mathcal{M}^-(g^-_\tau, g^-_N) &= 0,\\ \gamma^-\mathcal{M}^-(g^-_\tau, g^-_N) &= \left(g^-_\tau - \gamma^-\mathcal{L}_\tau(0, g^-_N), 0\right),\\ \mathcal{M}^-(g^-_\tau, g^-_N)|_{t=0} &= 0.\end{aligned}$$

One can analogously define the operator $\mathcal{M}^+$:

$$\begin{aligned}\mathcal{M}^+_\tau(g^+_\tau, g^+_N) &= U^+|\xi|F_1\left(g^+_\tau - \gamma^+\mathcal{L}_\tau(g^+_N, 0), 0\right) + F_1\left(g^+_\tau - \gamma^+\mathcal{L}_\tau(g^+_N, 0), 0\right),\\ \mathcal{M}^+_N(g^+_\tau, g^+_N) &= U^+|\xi|N'F_1\left(g^+_\tau - \gamma^+\mathcal{L}_\tau(g^+_N, 0), 0\right).\end{aligned}$$

It has the property of being divergence-free and of solving Oseen equations with zero normal boundary condition at $y = 1$, namely,

$$\begin{aligned}(\partial_t - \nu\Delta + U\partial_x)\mathcal{M}^+(g_\tau^+, g_N^+) + \boldsymbol{\nabla} p &= 0,\\ \boldsymbol{\nabla}\cdot\mathcal{M}^+(g_\tau^+, g_N^+) &= 0,\\ \gamma^+\mathcal{M}^+(g_\tau^+, g_N^+) &= \left(g_\tau^+ - \gamma^+\mathcal{L}_\tau(g_N^+, 0), 0\right),\\ \mathcal{M}^-(g_\tau^+, g_N^+)|_{t=0} &= 0.\end{aligned}$$

4.4. The half space Oseen operators. One can finally define the half space Oseen operators $\mathcal{O}^+$ and $\mathcal{O}^-$. These operators solve the Oseen equations, are divergence-free, and have the right boundary condition at $y = +1$ and $y = -1$, respectively. Moreover, the trace of these operators, evaluated at $y = -1$ and $y = +1$, respectively, has norm strictly less than the boundary datum. This property, expressed in Propositions 4.5 and 4.6 below, makes them suitable for the iterative procedure of the next section.

The operator $\mathcal{O}^+$ is defined as

$$\begin{aligned}\mathcal{O}_N^+(g_\tau^+, g_N^+) &= \mathcal{L}_N(g_N^+, 0) + \mathcal{M}_N^+(g_\tau^+, g_N^+)\ ,\\ \mathcal{O}_\tau^+(g_\tau^+, g_N^+) &= \mathcal{L}_\tau(g_N^+, 0) + \mathcal{M}_\tau^+(g_\tau^+, g_N^+)\ .\end{aligned}$$

The operator $\mathcal{O}^-$ is defined as

$$\begin{aligned}\mathcal{O}_\tau^-(g_\tau^-, g_N^-) &= \mathcal{L}_\tau(0, g_N^-) + \mathcal{M}_\tau^-(g_\tau^-, g_N^-)\ ,\\ \mathcal{O}_N^-(g_\tau^-, g_N^-) &= \mathcal{L}_N(0, g_N^-) + \mathcal{M}_N^-(g_\tau^-, g_N^-)\ .\end{aligned}$$

These operators have the property of solving the Oseen equations in the channel: $\mathcal{O}^+$ with the *right* boundary condition at $y = +1$; $\mathcal{O}^-$ with the *right* boundary condition at $y = -1$.

$$\begin{aligned}(\partial_t - \nu\Delta + U\partial_x)\mathcal{O}^\pm(g_\tau^\pm, g_N^\pm) + \boldsymbol{\nabla} p &= 0,\\ \boldsymbol{\nabla}\cdot\mathcal{O}^\pm(g_\tau^\pm, g_N^\pm)\ &= 0,\\ \gamma^\pm\mathcal{O}^\pm(g_\tau^\pm, g_N^\pm) &= (g_\tau^\pm, g_N^\pm),\\ \mathcal{O}^\pm(g_\tau^-, g_N^-)|_{t=0} &= 0\ .\end{aligned}$$

We now give some estimates on these operators. In the next section these estimates will allow us to construct the solution of the Oseen equation in the channel. In these estimates we shall always suppose that the normal component of the datum is such that $|\xi|^{-1}g_N \in H_T^{\prime l+1}$, i.e., that $\boldsymbol{g} \in \boldsymbol{H}_T^{\prime l}$.

PROPOSITION 4.7. *Let $\boldsymbol{g}^+ \in \boldsymbol{H}_T^{\prime l}$ with $\boldsymbol{g}^+|_{t=0} = 0$. Then $\mathcal{O}^+(\boldsymbol{g}^+) \in L_T^l$ and the following estimate holds:*

$$\|\mathcal{O}^+(\boldsymbol{g}^+)\|_{l,T} \leq c\ \boldsymbol{|}\,\boldsymbol{g}^+\,\boldsymbol{|}_{\,l,T}\ .$$

PROPOSITION 4.8. *Let $\boldsymbol{g}^- \in \boldsymbol{H}_T^{\prime l}$ with $\boldsymbol{g}^-|_{t=0} = 0$. Then $\mathcal{O}^-(\boldsymbol{g}^-) \in \boldsymbol{H}_T^l$ and the following estimate holds:*

$$\|\mathcal{O}^-(\boldsymbol{g}^-)\|_{l,T} \leq c\ \boldsymbol{|}\,\boldsymbol{g}^-\,\boldsymbol{|}_{\,l,T}\ .$$

The proof of these propositions can be easily achieved by using Propositions 4.4, 4.1, and 3.3.

We now give the estimates on the trace of the operators $\mathcal{O}^+$ and $\mathcal{O}^-$. We shall also prove that the trace at $y = -1$ (at $y = +1$, respectively) of $\mathcal{O}^+$ (of $\mathcal{O}^-$, respectively) has the integrability property for the normal influx.

Proposition 4.9. *Let $\boldsymbol{g}^+ \in \boldsymbol{H}'^{l}_{T}$ with $\boldsymbol{g}^+|_{t=0} = 0$. Then $\gamma^-\mathcal{O}^+ \in \boldsymbol{H}'^{l}_{T}$. Moreover, for any $0 < \alpha < 1$ there exists $T_\alpha > 0$ such that*

$$\left| \gamma^-\mathcal{O}^+(\boldsymbol{g}^+) \right|_{l,T_\alpha} \leq \alpha \left| \boldsymbol{g}^+ \right|_{l,T} . \tag{4.14}$$

Proposition 4.10. *Let $\boldsymbol{g}^- \in \boldsymbol{H}'^{l}_{T}$ with $\boldsymbol{g}^-|_{t=0} = 0$. Then $\gamma^+\mathcal{O}^-(\boldsymbol{g}^-) \in \boldsymbol{H}'^{l}_{T}$. Moreover, for any $0 < \alpha < 1$ there exists $T_\alpha > 0$ such that*

$$\left| \gamma^+\mathcal{O}^-(\boldsymbol{g}^-) \right|_{l,T_\alpha} \leq \alpha \left| \boldsymbol{g}^- \right|_{l,T} . \tag{4.15}$$

The proof of Propositions 4.9 and 4.10 is given in the appendix.

4.5. The Oseen operator with boundary data. We now have to solve the Oseen equations (4.1)–(4.5). We shall construct the solution as an infinite sum:

$$\boldsymbol{u} = \sum_{i=0}^{\infty} \boldsymbol{u}^{(i)} . \tag{4.16}$$

Each term of the series solves the Oseen equations. The zeroth term $\boldsymbol{u}^{(0)}$ has the right boundary condition at $y = -1$:

$$\boldsymbol{u}^{(0)} = \mathcal{O}^-(\boldsymbol{g}^-) . \tag{4.17}$$

As the boundary condition for the first term of the series we choose $\gamma^+\boldsymbol{u}^{(1)} = \boldsymbol{g}^+ - \gamma^+\boldsymbol{u}^{(0)}$. Therefore, one has the following expression for $\boldsymbol{u}^{(1)}$:

$$\boldsymbol{u}^{(1)} = \mathcal{O}^+\left(\boldsymbol{g}^+ - \gamma^+\boldsymbol{u}^{(0)}\right) . \tag{4.18}$$

In fact this choice will fix the boundary condition at $y = +1$; on the other end it generates a boundary datum at $y = -1$. We define the second term of the series $\boldsymbol{u}^{(2)}$ so that it cancels this boundary datum at $y = -1$:

$$\boldsymbol{u}^{(2)} = \mathcal{O}^-\left(-\gamma^-\boldsymbol{u}^{(1)}\right) . \tag{4.19}$$

Recursively we define the generic even and odd term of the series:

$$\boldsymbol{u}^{(2m)} = \mathcal{O}^-\left(-\gamma^-\boldsymbol{u}^{(2m-1)}\right) , \qquad m > 1 , \tag{4.20}$$

$$\boldsymbol{u}^{(2m+1)} = \mathcal{O}^+\left(-\gamma^+\boldsymbol{u}^{(2m)}\right) , \qquad m > 1 . \tag{4.21}$$

To prove the convergence of the series we prove the following proposition.

Proposition 4.11. *Let $\boldsymbol{g}^\pm \in \boldsymbol{H}'^{l}_{T}$ with $\boldsymbol{g}^\pm|_{t=0} = 0$. Then for any $0 < \alpha < 1$ there exists $T_\alpha > 0$ such that for each $\boldsymbol{u}^{(i)}$ the following estimate holds:*

$$\|\boldsymbol{u}^{(i)}\|_{l,T_\alpha} \leq c\alpha^{i-2}\left(\left| \boldsymbol{g}^+ \right|_{l,T} + \left| \boldsymbol{g}^- \right|_{l,T} \right) . \tag{4.22}$$

Proof. We prove the proposition when $i = 2m$. The proof is analogous when i is odd:

$$\begin{aligned}\boldsymbol{u}^{(2m)} &= \mathcal{O}^-(-\gamma^-\boldsymbol{u}^{(2m-1)})\\ &= \mathcal{O}^-(-\gamma^-\mathcal{O}^+(-\gamma^+\boldsymbol{u}^{(2m-2)}))\\ &= \cdots\\ &= \mathcal{O}^-\underbrace{(-\gamma^-\mathcal{O}^+(-\gamma^+\mathcal{O}^-(-\gamma^-\mathcal{O}^+(\cdots(-\gamma^-\mathcal{O}^+}_{2m-2\text{ times}}(\boldsymbol{g}^+-\gamma^+\mathcal{O}^-\boldsymbol{g}^-))\ldots)))).\end{aligned}$$

Hence, first using Proposition 4.8, then using Propositions 4.9 and 4.10, one has

$$\begin{aligned}\|\boldsymbol{u}^{(2m)}\|_{l,T_\alpha} &= \|\mathcal{O}^-(-\gamma^-\boldsymbol{u}^{(2m-1)})\|_{l,T_\alpha}\\ &\le c\,\left|\,\gamma^-\boldsymbol{u}^{(2m-1)}\,\right|_{l,T_\alpha} = c\,\left|\,\gamma^-\mathcal{O}^+(\gamma^+\boldsymbol{u}^{(2m-2)})\,\right|_{l,T_\alpha}\\ &\le c\alpha\,\left|\,\gamma^+\boldsymbol{u}^{(2m-2)}\,\right|_{l,T_\alpha}\\ &\le\cdots\\ &\le c\alpha^{2m-2}\left(\left|\,\boldsymbol{g}^+\,\right|_{l,T}+\left|\,\boldsymbol{g}^-\,\right|_{l,T}\right)\;.\end{aligned}$$

Therefore the series (4.16) with $\boldsymbol{u}^{(i)}$ given by (4.17)–(4.21) is convergent, and we can define the operator $\mathcal{O}_b$, solving Oseen equations with boundary data:

$$\mathcal{O}_b(\boldsymbol{g}^-,\boldsymbol{g}^+) = \sum_{i=0}^{\infty}\boldsymbol{u}^{(i)}\;. \tag{4.23}$$

The following theorem is the main result of this section.

THEOREM 4.12. *Let* $\boldsymbol{g}^\pm = (g_\tau^\pm, g_N^\pm)$ *such that* $\boldsymbol{g}_N^\pm \in \boldsymbol{H}_T'^{l}$ *and* $\boldsymbol{g}^\pm|_{t=0} = 0$. *Then there exists* T_α, *independent of* $\boldsymbol{g}^\pm$, *such that the operator* $\mathcal{O}_b$, *defined by* (4.23) *with* $\boldsymbol{u}^{(i)}$ *given by* (4.17)–(4.21), *represents the solutions of the Oseen equations* (4.1)–(4.5). *Moreover,* $\mathcal{O}_b(\boldsymbol{g}^-,\boldsymbol{g}^+) \in L^l_{T_\alpha}$ *and the following estimate holds:*

$$\|\mathcal{O}_b(\boldsymbol{g}^-,\boldsymbol{g}^+)\|_{l,T_\alpha} \le c\left[\left|\,\boldsymbol{g}^+\,\right|_{l,T}+\left|\,\boldsymbol{g}^-\,\right|_{l,T}\right]\;.$$

5. The explicit solution of the Oseen equations.

We now solve the Oseen equations with source term, boundary data, and initial data (1.1)–(1.5). To accomplish this task first we introduce the projection operator onto the divergence-free part of a vector function. We shall write this projection operator so that its normal component evaluated at the boundary is identically zero. Then we shall use this projection operator to project the convection-diffusion operator F_0 and F_2 to get the operators $\mathcal{O}_i$ and $\mathcal{O}_s$. These operators solve the Oseen equation with initial datum and source term, respectively. On the other hand, they generate *wrong* boundary data. Finally, in the last subsection we shall use the operator $\mathcal{O}_b$ to cancel the wrong boundary data and get the representation of the solution of (1.1)–(1.5).

5.1. The projection operator.

We introduce the following projection operator:

$$P = 1 - \boldsymbol{\nabla}\Delta_N^{-1}\boldsymbol{\nabla}\cdot\;\;. \tag{5.1}$$

In the above expression with Δ_N^{-1} we have denoted the operator that solves the Poisson equation with Neumann boundary conditions. We give the explicit expression of this projection operator:

$$\begin{aligned}
P_\tau\ \boldsymbol{w} = w_\tau &+ \frac{N'}{\sinh(2|\xi'|)}\left[\gamma^+ w_N \cosh[|\xi'|(y+1)] - \gamma^- w_N \cosh[|\xi'|(y-1)]\right] \\
&- \frac{|\xi'|}{2\sinh(2|\xi'|)}\left\{\int_{-1}^{y} dy' \left[\cosh(|\xi'|(y-y'-2)) + \cosh(|\xi'|(y+y'))\right] w_\tau(|\xi'|, y')\right. \\
&- \int_{-1}^{y} dy' \left[\sinh(|\xi'|(y-y'-2)) - \sinh(|\xi'|(y+y'))\right] N' w_N(|\xi'|, y') \\
&+ \int_{y}^{1} dy' \left[\cosh(|\xi'|(y-y'+2)) + \cosh(|\xi'|(y+y'))\right] w_\tau(|\xi'|, y') \\
&\left. - \int_{y}^{1} dy' \left[\sinh(|\xi'|(y-y'+2)) - \sinh(|\xi'|(y+y'))\right] N' w_N(|\xi'|, y')\right\}, \\
P_N \boldsymbol{w} = &\frac{1}{\sinh(2|\xi'|)}\left\{\gamma^+ w_N \sinh[|\xi'|(y+1)] - \gamma^- w_N \sinh[|\xi'|(y-1)]\right\} \\
&+ \frac{|\xi'|}{2\sinh(2|\xi'|)}\left\{\int_{-1}^{y} dy'\ \left[\sinh(|\xi'|(y-y'-2)) + \sinh(|\xi'|(y+y'))\right]\ N' w_\tau\right. \\
&+ \int_{-1}^{y} dy'\ \left[\cosh(|\xi'|(y-y'-2)) - \cosh(|\xi'|(y+y'))\right]\ w_N \\
&+ \int_{y}^{1} dy'\ \left[\sinh(|\xi'|(y-y'+2)) + \sinh(|\xi'|(y+y'))\right]\ N' w_\tau \\
&\left. + \int_{y}^{1} dy'\ \left[\cosh(|\xi'|(y-y'+2)) - \cosh(|\xi'|(y+y'))\right]\ w_N\right\}.
\end{aligned}$$

It is not difficult to see that the following proposition holds.

Proposition 5.1. *Let $\boldsymbol{w} \in L^l_T$. Then $P\boldsymbol{w} \in L^l_T$ and*

$$\|P\boldsymbol{w}\|_{l,T} \le c\,\|\boldsymbol{w}\|_{l,T}\ .$$

It is important to notice that, if $\gamma^\pm w_N = 0$, then the normal component of the projection operator evaluated at the boundary is identically zero:

$$\text{if} \quad \gamma^\pm w_N = 0, \qquad \text{then} \quad \gamma^\pm P_N \boldsymbol{w} = 0\ . \tag{5.2}$$

5.2. The projected convection-diffusion operators. We can now introduce the operators $\mathcal{O}_i$ and $\mathcal{O}_s$, defined as

$$\mathcal{O}_i = PF_0\ , \qquad \mathcal{O}_s = PF_2\ .$$

Supposing that $\boldsymbol{\nabla}\cdot\boldsymbol{u}_0 = 0$, one therefore has that $\mathcal{O}_i\boldsymbol{u}_0$ satisfies

$$\begin{aligned}
\left(\partial_t - \partial_{YY} - \varepsilon^2\partial_{xx}\right)\mathcal{O}_i\boldsymbol{u}_0 + \boldsymbol{\nabla}p &= 0, \\
\boldsymbol{\nabla}\cdot\mathcal{O}_i\boldsymbol{u}_0 &= 0, \\
\mathcal{O}_i\boldsymbol{u}_0|_{t=0} &= \boldsymbol{u}_0\ .
\end{aligned}$$

On the other hand, $\mathcal{O}_s\boldsymbol{w}$ satisfies

$$\begin{aligned}
\left(\partial_t - \partial_{YY} - \varepsilon^2\partial_{xx}\right)\mathcal{O}_s\boldsymbol{w} + \boldsymbol{\nabla}p &= \boldsymbol{w}, \\
\boldsymbol{\nabla}\cdot\mathcal{O}_s\boldsymbol{w} &= 0, \\
\mathcal{O}_s\boldsymbol{w}|_{t=0} &= 0\ .
\end{aligned}$$

The following estimates are a consequence of the properties of P, expressed in Proposition 5.1, and of the properties of F_0 and F_2, expressed in Propositions 3.1 and 3.4.

PROPOSITION 5.2. *Let $\boldsymbol{u}_0 \in L^l$. Then $\mathcal{O}_i \boldsymbol{u}_0 \in L^l_T$ and*

$$\|\mathcal{O}_i \boldsymbol{u}_0\|_{l,T} \le \|\boldsymbol{u}_0\|_l \, .$$

PROPOSITION 5.3. *Let $\boldsymbol{w} \in L^l_T$. Then $\mathcal{O}_s \boldsymbol{w} \in L^l_T$ and*

$$\|\mathcal{O}_s \boldsymbol{w}\|_{l,T} \le c \, \|\boldsymbol{w}\|_{l,T} \, .$$

The normal components of the operators $\mathcal{O}_i$ and $\mathcal{O}_s$ evaluated at the boundary are zero (because of (5.2)). This readily gives the following estimates on the traces of $\mathcal{O}_i$ and $\mathcal{O}_s$.

PROPOSITION 5.4. *Let $\boldsymbol{u}_0 \in L^l$. Then $\gamma^\pm \mathcal{O}_i \boldsymbol{u}_0 \in \boldsymbol{H}'^l_T$ and*

$$\bigl|\, \gamma^\pm \mathcal{O}_i \boldsymbol{u}_0 \,\bigr|_{l,T} \le c \|\boldsymbol{u}_0\|_l \, .$$

PROPOSITION 5.5. *Let $\boldsymbol{w} \in L^l_T$. Then $\gamma^\pm \mathcal{O}_s \boldsymbol{w} \in \boldsymbol{H}'^l_T$ and*

$$\bigl|\, \gamma^\pm \mathcal{O}_s \boldsymbol{w} \,\bigr|_{l,T} \le c \|\boldsymbol{w}\|_{l,T} \, .$$

5.3. The solution of the Oseen equations. We can finally introduce the operator $\mathcal{O}$ that solves (1.1)–(1.5):

$$\begin{aligned} &\mathcal{O}(\boldsymbol{f}, \boldsymbol{u}_0, \boldsymbol{g}^-, \boldsymbol{g}^+) \\ &\quad = \mathcal{O}_s \boldsymbol{f} + \mathcal{O}_i \boldsymbol{u}_0 + \mathcal{O}_b(\boldsymbol{g}^+ - \gamma^+ \mathcal{O}_s \boldsymbol{f} - \gamma^+ \mathcal{O}_i \boldsymbol{u}_0, \boldsymbol{g}^- - \gamma^- \mathcal{O}_s \boldsymbol{f} - \gamma^- \mathcal{O}_i \boldsymbol{u}_0) \, . \end{aligned} \tag{5.3}$$

Therefore, if one defines

$$\boldsymbol{u} = \mathcal{O}(\boldsymbol{f}, \boldsymbol{u}_0, \boldsymbol{g}^+, \boldsymbol{g}^-) \, , \tag{5.4}$$

then one has that $\boldsymbol{u}$ solves the system (1.1)–(1.5).

Remark 5.1. The representation of the solution given by the operator $\mathcal{O}$ is valid only up to the time T_α. In fact the series which defines the operator $\mathcal{O}_b$ converges only up to the time T_α. On the other hand, given that T_α does not depend on the boundary data, one can take the value of the solution at the time T_α as the initial datum and solve the corresponding Oseen problem up to the time $2T_\alpha$. One can therefore construct the solution up to the time T. Moreover, we also notice that the time $T_\alpha = O(\nu^{-1})$. Therefore, in the zero viscosity limit, the representation (5.4) is valid for an arbitrarily long time. The following theorem holds.

THEOREM 5.6. *Suppose $\boldsymbol{f} \in L^l_T$, $\boldsymbol{u}_0 \in L^l$, and $\boldsymbol{g}^\pm \in \boldsymbol{H}'^l_T$. Suppose the following compatibility conditions are verified:*

$$\boldsymbol{\nabla} \cdot \boldsymbol{u}_0 = 0 \, , \tag{5.5}$$

and

$$\boldsymbol{g}^\pm|_{t=0} = \gamma^\pm \boldsymbol{u}_0 \, . \tag{5.6}$$

Then the solution of the Oseen equations (1.1)–(1.5) *is represented, for a time T_α, by* (5.4). *Moreover, the following estimate holds:*

$$\|\mathcal{O}(\boldsymbol{f}, \boldsymbol{u}_0, \boldsymbol{g}^-, \boldsymbol{g}^+)\|_{l,T} \le c \Bigl[\|\boldsymbol{f}\|_{l,T} + \|\boldsymbol{u}_0\|_l + \bigl|\, \boldsymbol{g}^+ \,\bigr|_{l,T} + \bigl|\, \boldsymbol{g}^- \,\bigr|_{l,T} \Bigr] \, .$$

The proof of this theorem is a consequence of Propositions 5.2, 5.3, 5.4, and 5.5 and Theorem 4.12.

We finally give a proposition which will be useful in the solution of the error equation, given in the next section. It says that if the forcing term in (1.1) is of the same order of the square root of the viscosity, then one can use the heat kernel to increase the regularity of the solution.

PROPOSITION 5.7. *Suppose* $\boldsymbol{f} = \sqrt{\nu}\boldsymbol{h}$ *with* $\boldsymbol{h} \in L_T^{l-1}$, $\boldsymbol{g}^{\pm} \in \boldsymbol{H}_T'^{l}$, *and* $\boldsymbol{u}_0 \in L'^{l}$. *Suppose that the compatibility conditions* (5.5) *and* (5.6) *are satisfied. Then* $\mathcal{O}_s(\boldsymbol{f}, \boldsymbol{g}^-, \boldsymbol{g}^+) \in L_T^l$, *and the following estimate holds:*

$$\|\mathcal{O}(\boldsymbol{f}, \boldsymbol{u}_0, \boldsymbol{g}^+, \boldsymbol{g}^-)\|_{l,T} \leq c\left[\|\boldsymbol{h}\|_{l-1,T} + \|\boldsymbol{u}_0\|_l + \left|\!\left|\!\left| \boldsymbol{g}^+ \right|\!\right|\!\right|_{l,T} + \left|\!\left|\!\left| \boldsymbol{g}^- \right|\!\right|\!\right|_{l,T}\right].$$

The proof of this proposition can be easily achieved using the analogous estimate for the operator F_2 given in Proposition 3.5, and the estimates for P, $\mathcal{O}_s$, $\mathcal{O}_i$, and $\mathcal{O}_b$, given in Propositions 5.1, 5.2, 5.3, 5.4, and 5.5 and Theorem 4.12.

6. The asymptotic analysis. We are now ready to introduce the boundary layer analysis for the Oseen equations (1.1)–(1.5). For simplicity we shall suppose that the source term is not present. In a remark at the end of section 6.1 we shall sketch the procedure that allows us to handle a source term.

We shall impose the initial condition $\boldsymbol{u}_0 \in H^l$ and show that the solution is the sum of an inviscid (Euler) part, two boundary layers (Prandtl) parts exponentially decaying outside a region of size $\varepsilon = \sqrt{\nu}$ close to the two boundaries $y = -1$ and $y = +1$, and a correction term.

We seek a solution of the form

$$u = u^E + u^P + \varepsilon w_\tau, \tag{6.1}$$

$$v = v^E + \varepsilon v^P + \varepsilon w_N, \tag{6.2}$$

$$p = p^E + \varepsilon p^w\ , \tag{6.3}$$

where (u^E, v^E) represents the inviscid solution, $(u^P, \varepsilon v^P)$ represents the Prandtl part which describes the behavior of the fluid in the boundary layers close to $y = \pm 1$, and (w_τ, w_N) is the correction term. They solve the following equations.

1. The convective equations

$$(\partial_t + U\partial_x)\boldsymbol{u}^E + \boldsymbol{\nabla} p^E = 0, \tag{6.4}$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{u}^E = 0, \tag{6.5}$$

$$\gamma^- v^E = g_N^-, \tag{6.6}$$

$$\gamma^+ v^E = g_N^+, \tag{6.7}$$

$$\boldsymbol{u}^E(t = 0) = \boldsymbol{u}_0\ . \tag{6.8}$$

2. The boundary layer equations

$$(\partial_t - \varepsilon^2\Delta + U\partial_x)u^P = 0, \tag{6.9}$$

$$\partial_x u^P + \partial_Y v^P = 0, \tag{6.10}$$

$$\gamma^- u^P = g_\tau^- - \gamma^- u^E, \tag{6.11}$$

$$\gamma^+ u^P = g_\tau^+ - \gamma^+ u^E, \tag{6.12}$$

$$u^P(t = 0) = 0\ . \tag{6.13}$$

3. The correction equations

$$(\partial_t - \varepsilon^2\Delta + U\partial_x)\boldsymbol{w} + \boldsymbol{\nabla} p^w = \varepsilon\Delta \boldsymbol{u}^E, \tag{6.14}$$
$$\boldsymbol{\nabla}\cdot\boldsymbol{w} = 0, \tag{6.15}$$
$$\gamma^-\boldsymbol{w} = (0,\, -\gamma^- v^P\ ), \tag{6.16}$$
$$\gamma^+\boldsymbol{w} = (0,\, -\gamma^+ v^P\ ), \tag{6.17}$$
$$\boldsymbol{w}(t=0) = 0\ . \tag{6.18}$$

We require to hold the compatibility conditions between the boundary and the initial data:

$$\gamma^\pm v_0 = g_N^\pm(x, t=0), \tag{6.19}$$
$$g_\tau^\pm(t=0) = -\gamma^\pm u_0. \tag{6.20}$$

We require to hold the incompressibility condition for the initial data:

$$\boldsymbol{\nabla}\cdot\boldsymbol{u}_0 = 0. \tag{6.21}$$

Moreover, the incompressibility condition requires that

$$\int_{-\infty}^{\infty} (g_N^+ - g_N^-)\, dx = 0.$$

Notice the absence of the pressure term in (6.9)–(6.13). This is due to the introduction of a boundary layer corrector which differs from the usual Prandtl velocity for the value of the Euler velocity at the boundary. Namely, $\tilde{\boldsymbol{u}}^P = \boldsymbol{u}^P - \gamma\boldsymbol{u}^E$. Then, using the Euler equation at the boundary, one gets (6.9). For more details, see [2].

We now solve the above equations.

6.1. The Euler equations. One can see that the solution of (6.4)–(6.8) is

$$\begin{aligned} v^E &= \mathcal{L}_N\left(g_N^+ - v_0^E(x-Ut,1), g_N^- - v_0^E(x-Ut,-1)\right) \\ &\quad + v_0(x-Ut,y), \end{aligned} \tag{6.22}$$

$$\begin{aligned} u^E &= \mathcal{L}_\tau\left(g_N^+ - v_0^E(x-Ut,1), g_N^- - v_0^E(x-Ut,-1)\right) \\ &\quad + u_0(x-Ut,y), \end{aligned} \tag{6.23}$$

where the operator $\mathcal{L}$ was introduced in section 4.2. We now give an estimate on the above solution.

PROPOSITION 6.1. *Let* $\boldsymbol{g} \in \boldsymbol{H}'^l_T$. *Moreover, let* $\boldsymbol{u}_0 \in H^l$ *satisfy the compatibility conditions* (6.19) *and* (6.20) *and the incompressibility condition* (6.21). *Then* $\boldsymbol{u}^E \in H^l_T$ *and the following estimate holds:*

$$|\boldsymbol{u}^E|_{l,T} \le c(\,\big|\,\boldsymbol{g}^-\,\big|_{l,T} + \big|\,\boldsymbol{g}^+\,\big|_{l,T} + |\boldsymbol{u}_0|_l).$$

The proof is based on Proposition 4.4.

Remark 6.1. If a source term $\boldsymbol{f}$ is present in (6.4) one can solve the inviscid equation using the integrated (with respect to time) projection operator onto the divergence-free part.

6.2. The Prandtl equations. Equations (6.9) and the boundary conditions (6.11) and (6.12) and the initial condition (6.13) are convection-diffusion equations with boundary data and zero initial datum and no source. We have already solved these equations through the operator F_1 introduced in section 3. The solution is, therefore,

$$u^P = F_1(g_\tau^+ - \gamma^+ u^E, g_\tau^- - \gamma^- u^E) \; . \tag{6.24}$$

The normal component can be found using the incompressibility condition (6.10):

$$v^P = \int_Y^0 dY' \partial_x u^P \; . \tag{6.25}$$

Therefore, we conclude with the following proposition.

Proposition 6.2. *Let* $\mathbf{g}^\pm \in \boldsymbol{H}'^l_T$ *satisfy the compatibility conditions* (6.19)–(6.20) *and the incompressibility condition* (6.21). *Then* $u^P \in K_T^{l,\mu}$, $v^P \in K_T^{l-1,\mu}$, *and the following estimates hold:*

$$|u^P|_{l,\mu,T} \le c \left(\boldsymbol{|}\, \mathbf{g}^- \,\boldsymbol{|}_{l,T} + \boldsymbol{|}\, \mathbf{g}^+ \,\boldsymbol{|}_{l,T} + |\boldsymbol{u}_0|_l \right) ,$$
$$|v^P|_{l-1,\mu,T} \le c \left(\boldsymbol{|}\, \mathbf{g}^- \,\boldsymbol{|}_{l,T} + \boldsymbol{|}\, \mathbf{g}^+ \,\boldsymbol{|}_{l,T} + |\boldsymbol{u}_0|_l \right) .$$

6.3. The error equation. Let us now consider the correction term $\boldsymbol{w}$ satisfying (6.14)–(6.18).

One can see that (6.14)–(6.18) are of the same form as (1.1)–(1.5), namely, they are the Oseen equations with source term and boundary and initial data satisfying the hypotheses of Proposition 5.7. In fact the source term in (6.14) is $\varepsilon \Delta e^{i\xi U t} \boldsymbol{u}_0$, which is of the form $\varepsilon \boldsymbol{h}$ with $\boldsymbol{h} \in H_T^{l-2}$. Moreover, the boundary data $-\gamma^\pm v^P$ are of the form

$$-\gamma^\pm v^P = \partial_x \beta^\pm = -|\xi| N' \beta^\pm, \qquad \text{where} \qquad \beta^\pm = \int_0^{\pm 1/\varepsilon} dY' \, u^{P(\pm)} \in H'^l_T \; .$$

Therefore, $\gamma^\pm \boldsymbol{w} \in \boldsymbol{H}'^{l-1}_T$.

Hence the solution has the form given by (5.4) and we can give the estimate on the error $\boldsymbol{w}$.

Proposition 6.3. *Let us suppose that* $\boldsymbol{g}^\pm \in \boldsymbol{H}'^l_T$ *and* $\boldsymbol{u}_0 \in H^l$, *satisfying the compatibility conditions* (6.19)–(6.20) *and the incompressibility condition* (6.21). *Then the solution of* (6.14)–(6.18) $\boldsymbol{w} \in L_T^{l-1}$, *and the following estimate hold:*

$$\|w\|_{l-1,T} \le c \left[\boldsymbol{|}\, \boldsymbol{g}^+ \,\boldsymbol{|}_{l,T} + \boldsymbol{|}\, \boldsymbol{g}^- \,\boldsymbol{|}_{l,T} + |\boldsymbol{u}_0|_l \right] \; .$$

The above estimate is not enough to get the convergence of the solution of the Oseen equations to $\boldsymbol{u}^E + \boldsymbol{u}^P$ in a space where first derivatives are considered. A more refined analysis of the structure of the error is needed.

6.4. The structure of the error. We divide the error in the following way:

$$\boldsymbol{w} = \boldsymbol{w}^E + \boldsymbol{w}^{BL} + \varepsilon \boldsymbol{e}. \tag{6.26}$$

The Eulerian part of the error $\boldsymbol{w}^E$ satisfies convective equations, of the type (6.4)–(6.8), with $\varepsilon \Delta u^E$ as source term and with prescribed normal boundary data $-\gamma^\pm v^P$.

The boundary layer part of the error $\boldsymbol{w}^{BL}$ satisfies the boundary layer equations, of the type (6.9)–(6.13), with prescribed tangential component $-\gamma^{\pm}\boldsymbol{w}^{E}$.

Finally the overall correction $\boldsymbol{e}$ satisfies the correction equations, of the type (6.14)–(6.18), with zero source term and with prescribed boundary data $(0, -\gamma^{\pm} w_N^{BL})$, where with w_N^{BL} we have denoted the normal component of $\boldsymbol{w}^{BL}$.

One can therefore state the following proposition.

PROPOSITION 6.4. *Suppose the hypotheses of Proposition* 6.3 *hold true. Then the correction* $\boldsymbol{w}$ *admits a decomposition of the form* (6.26) *with* $\boldsymbol{w}^E \in H_T^{l-1}$, $w_\tau^{BL} \in K_T^{l-1,\mu}$, $w_N^{BL} \in K_T^{l-2,\mu}$, $\boldsymbol{e} \in L_T^{l-2}$, *and the following estimate holds:*

$$|\boldsymbol{w}^E|_{l-1,T} + |w_\tau^{BL}|_{l-1,\mu,T} + |w_N^{BL}|_{l-2,\mu,T} + \|\boldsymbol{e}\|_{l-2,T} \leq c\Big(|\!|\!| \boldsymbol{g}^+ |\!|\!|_{l,T} + |\!|\!| \boldsymbol{g}^- |\!|\!|_{l,T} + |\boldsymbol{u}_0|_l \Big) . \tag{6.27}$$

6.5. The main results. We can summarize the results of this section in the following theorem which is the main result of this paper.

THEOREM 6.5. *Let* $\boldsymbol{g}^{\pm} \in \boldsymbol{H}_T^{\prime l}$ *and* $\boldsymbol{u}_0 \in H^l$ *satisfy the compatibility conditions* (6.19)–(6.20) *and the incompressibility condition* (6.21). *Then the solution of Oseen equations* (1.1)–(1.5) *can be written in the form*

$$\boldsymbol{u} = \boldsymbol{u}^E + \boldsymbol{u}^P + \varepsilon \boldsymbol{w} \ , \tag{6.28}$$

where $\boldsymbol{u}^E$ *satisfies* (6.4)–(6.8), $\boldsymbol{u}^P$ *satisfies* (6.9)–(6.13), *and* $\boldsymbol{w}$ *satisfies* (6.14)–(6.18). *Moreover,* $\boldsymbol{u}^E \in H_T^l$, $u^P \in K_T^{l,\mu}$, $v^P \in K_T^{l-1,\mu}$, *and* $\boldsymbol{w} \in L_T^{l-1}$. *Moreover, the correction can be decomposed as*

$$\boldsymbol{w} = \boldsymbol{w}^E + \boldsymbol{w}^{BL} + \varepsilon \boldsymbol{e} \tag{6.29}$$

with $\boldsymbol{w}^E \in H_T^{l-1}$, $w_\tau^{BL} \in K_T^{l-1,\mu}$, $w_N^{BL} \in K_T^{l-2,\mu}$, *and* $\boldsymbol{e} \in L_T^{l-2}$. *The following estimates hold:*

$$\begin{aligned}
|\boldsymbol{u}^E|_{l,T} &\leq c\Big(|\!|\!| \boldsymbol{g}^+ |\!|\!|_{l,T} + |\!|\!| \boldsymbol{g}^- |\!|\!|_{l,T} + |\boldsymbol{u}_0|_l \Big) , \\
|u^P|_{l,\mu,T} &\leq c\Big(|\!|\!| \boldsymbol{g}^+ |\!|\!|_{l,T} + |\!|\!| \boldsymbol{g}^- |\!|\!|_{l,T} + |\boldsymbol{u}_0|_l \Big) , \\
|v^P|_{l-1,\mu,T} &\leq c\Big(|\!|\!| \boldsymbol{g}^+ |\!|\!|_{l,T} + |\!|\!| \boldsymbol{g}^- |\!|\!|_{l,T} + |\boldsymbol{u}_0|_l \Big) , \\
|\boldsymbol{w}^E|_{l-1,T} &\leq c\Big(|\!|\!| \boldsymbol{g}^+ |\!|\!|_{l,T} + |\!|\!| \boldsymbol{g}^- |\!|\!|_{l,T} + |\boldsymbol{u}_0|_l \Big) , \\
|w_\tau^{BL}|_{l-1,\mu,T} &\leq c\Big(|\!|\!| \boldsymbol{g}^+ |\!|\!|_{l,T} + |\!|\!| \boldsymbol{g}^- |\!|\!|_{l,T} + |\boldsymbol{u}_0|_l \Big) , \\
|w_N^{BL}|_{l-2,\mu,T} &\leq c\Big(|\!|\!| \boldsymbol{g}^+ |\!|\!|_{l,T} + |\!|\!| \boldsymbol{g}^- |\!|\!|_{l,T} + |\boldsymbol{u}_0|_l \Big) , \\
\|\boldsymbol{e}\|_{l-2,T} &\leq c\Big(|\!|\!| \boldsymbol{g}^+ |\!|\!|_{l,T} + |\!|\!| \boldsymbol{g}^- |\!|\!|_{l,T} + |\boldsymbol{u}_0|_l \Big) .
\end{aligned}$$

From the above theorem the following estimate on the convergence of the Oseen equation to $\boldsymbol{u}^E + \boldsymbol{u}^P$ easily follows.

COROLLARY 6.1. *Let* $\boldsymbol{u}_0 \in H^4$ *and* $\boldsymbol{g}^{\pm} \in L^{\infty}([0,T], H^4)$ *satisfy the compatibility conditions* (6.19)–(6.20) *and the incompressibility condition* (6.21). *Moreover, suppose that the inflows at the boundaries* $g_N^{\pm}$ *are such that* $|\xi|^{-1} g_N^{\pm} \in L^{\infty}\big([0,T], L^2\big)$. *If* $\boldsymbol{u}$, $\boldsymbol{u}^E$, *and* $\boldsymbol{u}^P$ *denote the solutions of* (1.1)–(1.5), (6.4)–(6.8), *and* (6.9)–(6.13), *respectively, then one has the following estimate:*

$$\|\boldsymbol{u} - (\boldsymbol{u}^E + \boldsymbol{u}^P)\|_{L^{\infty}([0,T],H^2)} \leq c\varepsilon^{1/2} \left[\|\boldsymbol{u}_0\|_{H^4} + \|\boldsymbol{g}^+\|_{L^{\infty}([0,T],H^4)} + \|\boldsymbol{g}^-\|_{L^{\infty}([0,T],H^4)} \right] . \tag{6.30}$$

The factor $\varepsilon^{1/2}$ comes from the fact that the correction $\boldsymbol{w}$ also has a boundary layer structure, i.e., its derivative with respect to y is $O(\varepsilon^{-1})$ in a region of size $O(\varepsilon^{1/2})$.

Remark 6.2. With minor formal modifications in the definitions of the function spaces H^l given in section 2, which allow l to be equal to 1, Corollary 6.1 could be stated for initial data belonging to H^3.

Appendix.

Proof of Lemma 4.2. Let us consider first the L^2 norm.

$$
\begin{aligned}
&\sup_{0\le t\le T} \|\gamma^- U^+ F_1(g_\tau^+,0)\|^2_{L^2(\xi')} \\
&= \sup_{0\le t\le T}\left\| \int_{-1}^{1} dy' e^{-|\xi'|(1+y')} \int_0^t ds\, e^{-[\nu\xi'^2 + i\xi' U](t-s)} \sum_{n=-\infty}^{\infty} \frac{y'-1-4n}{(t-s)} \frac{e^{-\frac{(y'-1-4n)^2}{4\nu(t-s)}}}{\sqrt{4\pi\nu(t-s)}} g_\tau^+(\xi',s) \right\|^2_{L^2(\xi')} \\
&\le \sup_{0\le t\le T}\left\| \int_0^t ds g_\tau^+ \sum_{n=0}^{\infty} e^{-|\xi'|(4n+2)} \int_{-1}^{1} dy' \frac{y'-1-4n}{(t-s)} \frac{e^{-\left[\frac{y'-1-4n}{\sqrt{4\nu(t-s)}} + \sqrt{\nu}|\xi'|(t-s)^{1/2}\right]^2}}{\sqrt{4\pi\nu(t-s)}} \right\|^2 \\
&\quad + \sup_{0\le t\le T}\left\| \int_0^t ds\, g_\tau^+ \sum_{n=-\infty}^{-1} \int_{-1}^{1} dy' e^{|\xi'|(-2y'+4n)} \frac{y'-1-4n}{(t-s)} \frac{e^{-\left[\frac{y'-1-4n}{\sqrt{4\nu(t-s)}} - \sqrt{\nu}|\xi'|(t-s)^{1/2}\right]^2}}{\sqrt{4\pi\nu(t-s)}} \right\|^2 \\
&\le \sup_{0\le t\le T}\left\| \int_0^t ds g_\tau^+ \sum_{n=0}^{\infty} e^{-|\xi'|(4n+2)} \int_{-1}^{1} dy' 2\sqrt{\nu} \frac{d}{dy'} \frac{e^{-\left[\frac{y'-1-4n}{\sqrt{4\nu(t-s)}} + \sqrt{\nu}|\xi'|(t-s)^{1/2}\right]^2}}{\sqrt{4\pi(t-s)}} \right\|^2 \\
&\quad + \sup_{0\le t\le T}\left\| \int_0^t ds g_\tau^+ \sum_{n=0}^{\infty} e^{-|\xi'|(4n+2)} \int_{-1}^{1} dy'\, 2\sqrt{\nu}|\xi'| \frac{e^{-\left[\frac{y'-1-4n}{\sqrt{4\nu(t-s)}} + \sqrt{\nu}|\xi'|(t-s)^{1/2}\right]^2}}{\sqrt{4\pi(t-s)}} \right\|^2 \\
&\quad + \sup_{0\le t\le T}\left\| \int_0^t ds g_\tau^+ \sum_{n=-\infty}^{-1} \int_{-1}^{1} dy' e^{|\xi'|(-2y'+4n)} 2\sqrt{\nu} \frac{d}{dy'} \frac{e^{-\left[\frac{y'-1-4n}{\sqrt{4\nu(t-s)}} - \sqrt{\nu}|\xi'|(t-s)^{1/2}\right]^2}}{\sqrt{4\pi(t-s)}} \right\|^2 \\
&\quad + \sup_{0\le t\le T}\left\| \int_0^t ds g_\tau^+ \sum_{n=-\infty}^{-1} \int_{-1}^{1} dy' e^{|\xi'|(-2y'+4n)} 2\sqrt{\nu}|\xi'| \frac{e^{-\left[\frac{y'-1-4n}{\sqrt{4\nu(t-s)}} - \sqrt{\nu}|\xi'|(t-s)^{1/2}\right]^2}}{\sqrt{4\pi(t-s)}} \right\|^2 \\
&= I_1 + I_2 + I_3 + I_4\ .
\end{aligned}
$$

We now consider the four terms separately. We begin from I_1:

$$
I_1 = \sup_{0\le t\le T}\left\| \int_0^t ds\, g_\tau^+ \sqrt{\nu} \sum_{n=0}^{\infty} e^{-|\xi'|(4n+2)} \right.
$$

$$\times \left[\frac{e^{-\left[\frac{4n}{\sqrt{4\nu(t-s)}}-\sqrt{\nu}|\xi'|(t-s)^{1/2}\right]^2}}{\sqrt{\pi(t-s)}} - \frac{e^{-\left[\frac{4n+2}{\sqrt{4\nu(t-s)}}-\sqrt{\nu}|\xi'|(t-s)^{1/2}\right]^2}}{\sqrt{\pi(t-s)}} \right] \Bigg\|^2$$

$$\leq \sup_{0\leq t\leq T} \left\| \int_0^t ds\ g_\tau^+ 2\sqrt{\nu} \sum_{n=0}^{\infty} e^{-|\xi'|(4n+2)} \frac{e^{-\left[\frac{4n}{\sqrt{4\nu(t-s)}}-\sqrt{\nu}|\xi'|(t-s)^{1/2}\right]^2}}{\sqrt{\pi(t-s)}} \right\|^2$$

$$\leq \sup_{0\leq t\leq T} \left\| \int_0^t ds\ 2\sqrt{\nu}\, e^{-\nu|\xi'^2|(t-s)}\, e^{-2|\xi|} g_\tau^+ \sum_{n=0}^{\infty} \frac{e^{-\frac{(4n)^2}{4\nu(t-s)}}}{\sqrt{\pi(t-s)}} \right\|^2$$

$$\leq \sup_{0\leq t\leq T} \left\| \int_0^t ds 4\sqrt{\nu} \frac{1}{\sqrt{\pi(t-s)}} g_\tau^+ \right\|^2$$

$$\leq \frac{16\nu T}{\pi} \sup_{0\leq t\leq T} \|g_\tau^+\|^2.$$

The estimate of the terms with the derivatives with respect to x and t is the same. Let us now pass to I_2. Introducing the variable $\eta = (y' - 1 - 4n)/\sqrt{4\nu(t-s)} + \sqrt{\nu}|\xi|(t-s)^{1/2}$, we get

$$I_2$$

$$= \sup_{0\leq t\leq T} \frac{1}{\pi} \left\| \int_0^t ds\, g_\tau^+(\xi', s)\, 4\nu|\xi'| \sum_{n=0}^{\infty} e^{-|\xi'|(4n+2)} \int_{\frac{-4n-2}{\sqrt{4\nu(t-s)}}+\sqrt{\nu}|\xi'|(t-s)^{1/2}}^{\frac{-4n}{\sqrt{4\nu(t-s)}}+\sqrt{\nu}|\xi'|(t-s)^{1/2}} d\eta\, e^{-\eta^2} \right\|^2$$

$$\leq \sup_{0\leq t\leq T} \left\| \int_0^t ds\, g_\tau^+\, 8\nu|\xi'| \left(e^{-2|\xi|} \frac{1}{2} T\nu|\xi| + \sum_{n>\frac{1}{2}T\nu|\xi|} e^{-|\xi'|(4n+2)} e^{-\left[\frac{4n}{\sqrt{4\nu(t-s)}}-\sqrt{\nu}|\xi'|(t-s)^{1/2}\right]} \right) \right\|^2$$

$$\leq \left(16\nu^4 T^4 + 4\nu^2 T^2\right) \sup_{0\leq t\leq T} \left\| g_\tau^+(\xi', s) \right\|^2.$$

The estimate of I_3 is analogous to the estimate of I_1, while the estimate of I_4 is analogous to the estimate of I_2. From the above estimates it is apparent that, choosing T_α small enough, one can make $\|\gamma^- U^+ F_1(g_\tau^+, 0)\|_{l.T_\alpha} < \alpha |g_\tau^+|_{l,T}$ with $\alpha < 1$. □

Remark A.1. It is interesting to notice that, in the zero viscosity limit, the time T_α up to which the above estimate is valid grows to infinity: $T_\alpha \sim \nu^{-1}$.

Proof of Proposition 4.9. The trace at $y = -1$ of the operator $\mathcal{O}^+$ is made of three terms: (1) the trace γ^- of $\mathcal{L}_\tau(\cdot, 0)$, which is estimated in Lemma 4.5; (2) the trace γ^- of $U^+ F_1(\cdot, 0)$, which is estimated in Lemma 4.2; (3) the trace γ^- of the operator $F_1(\cdot, 0)$, which can be estimated as follows:

$$\sup_{0\leq t\leq T} \left\| \gamma^- F_1(f_\tau^+, 0) \right\|_{L^2(\xi')}^2$$

$$= \sup_{0\leq t\leq T} \left\| \int_0^t ds\, e^{-[\nu\xi'^2 + i\xi' U](t-s)} \sum_{n=-\infty}^{\infty} \frac{2+4n}{\varepsilon(t-s)} \frac{e^{-\frac{(2+4n)^2}{4\nu(t-s)}}}{\sqrt{4\pi(t-s)}} g_\tau^+(\xi', s) \right\|_{L^2(\xi')}^2$$

$$
\begin{aligned}
&\leq \frac{4}{\pi} \sup_{0\leq t\leq T} \left\| \int_0^t ds\, e^{-\nu {\xi'}^2 (t-s)} \sum_{n=0}^{\infty} \frac{1+2n}{\varepsilon(t-s)} \frac{e^{-\frac{(1+2n)^2}{\nu(t-s)}}}{\sqrt{(t-s)}}\, g_\tau^+(\xi', s) \right\|^2_{L^2(\xi')} \\
&\leq \frac{\nu}{\pi} \sup_{0\leq t\leq T} \left\| \int_0^t ds\, \frac{e^{-\nu {\xi'}^2 (t-s)}}{(t-s)^{1/2}}\, g_\tau^+(\xi', s) \int_0^\infty d\eta\, \eta\, e^{-\eta^2} \right\|^2_{L^2(\xi')} \\
&\leq \frac{\nu T}{\pi} \sup_{0\leq t\leq T} \left\| g_\tau^+ \right\|^2_{L^2(\xi')} .
\end{aligned}
$$

As far as the term $\sup_{0\leq t\leq T} \|\gamma^- U^+ |\xi'| F_1(f_\tau^+ - \gamma^+ \mathcal{L}_\tau(f_N^+, 0), 0)\|^2_{L^2(\xi')}$ is concerned, the estimate is analogous to the one given in the proof of Lemma 4.2. The only things one has to use are the fact that $\sqrt{\nu(t-s)}|\xi'|e^{-\nu {\xi'}^2(t-s)}$ is bounded and the regularizing property of the integration with respect to time of the factor $\frac{1}{\sqrt{t-s}}$. The estimate in (4.14) is thus achieved. □

Proof of Proposition 4.10. The proof of Proposition 4.10 is analogous to the proof of Proposition 4.9. □

Acknowledgments. The authors thank Professor Giga and an anonymous referee for the useful comments that helped to improve the paper and the presentation of the results.

REFERENCES

[1] M. Sammartino, *The boundary layer analysis for Stokes equations on a half space*, Comm. Partial Differential Equations, 22 (1997), pp. 749–771.

[2] M. Sammartino and R.E. Caflisch , *Zero viscosity limit for analytic solutions of the Navier-Stokes equation on a half-space* I. *Existence for Euler and Prandtl equations*, Comm. Math. Phys., 192 (1998), pp. 433–461.

[3] M. Sammartino and R.E. Caflisch, *Zero viscosity limit for analytic solutions of the Navier-Stokes equation on a half-space* II. *Construction of the Navier-Stokes solution*, Comm. Math. Phys., 192 (1998), pp. 463–491.

[4] R. Temam and X. Wang, *Asymptotic analysis of the linearized Navier-Stokes equations in a channel*, Differential Integral Equations, 8 (1995), pp. 1591–1618.

[5] R. Temam and X. Wang, *Asymptotic analysis of Oseen type equations in a channel at small viscosity*, Indiana Univ. Math. J., 45 (1996), pp. 863–916.

[6] R. Temam and X. Wang, *Boundary layers for Oseen's type equation in space dimension three*, Russian J. Math. Phys., 5 (1998), pp. 227–246.

[7] S. Ukai, *A solution formula for the Stokes equation in* $\mathbb{R}^n_+$, Comm. Pure Appl. Math., 40 (1987), pp. 611–621.

[8] D.V. Widder, *The Heat Equation*, Academic Press, New York, 1975.

SIAM J. MATH. ANAL.
Vol. 33, No. 2, pp. 411–439

L^1 STABILITY FOR SYSTEMS OF CONSERVATION LAWS WITH A NONRESONANT MOVING SOURCE*

SEUNG-YEAL HA†

Abstract. In this paper, we study L^1 stability for systems of conservation laws with a moving source $u_t + f(u)_x = g(x - ct, u)$. The source is assumed to be nonresonant in that its speed c is different from the characteristic speeds of the system. We show that weak solutions are globally L^1 stable. Based on the modified Glimm scheme, we construct a robust nonlinear functional $H(t) = H[u(\cdot, t), v(\cdot, t)]$ which is equivalent to the L^1 distance of two solutions u, v and is nonincreasing in time t. This functional $H[u, v]$ consists of a linear part $L[u, v]$ measuring the L^1 distance, a quadratic part $Q_d[u, v]$ measuring nonlinear couplings between waves of different characteristic fields, a generalized entropy functional $E[u, v]$ capturing the nonlinearity of characteristic fields, and a new functional $Q_{so}[u, v]$ measuring the source effect on the L^1 distance.

Key words. conservation laws, L^1 stability, nonlinear functional

AMS subject classifications. 35L65, 35L45

PII. S0036141000373045

1. Introduction.

The purpose of this paper is to establish the L^1 stability of the initial value problem for systems of hyperbolic conservation laws with a moving source:

$$
\begin{aligned}
u_t + f(u)_x &= g(x - ct, u), \quad (x, t) \in R \times R_+, \\
u(x, 0) &= u_0(x), \quad x \in R,
\end{aligned} \tag{1.1}
$$

where $u \in \mathcal{N} \subset R^n$, $f : \mathcal{N} \to R^n$, and $g : R \times \mathcal{N} \to R^n$ denote the conserved quantities, the C^2 flux function, and the source, respectively. This system is assumed to be strictly hyperbolic. It is well known [15] that in general the system (1.1) does not admit classical solutions even for smooth initial data because of the nonlinearity of the flux function. Therefore, one needs to consider weak solutions.

DEFINITION 1.1. *A bounded measurable function $u(x, t)$ is a weak solution of (1.1) with given initial data $u_0(x)$ if and only if for $\phi \in C_c^1(R \times R_+)$,*

$$
\int_0^\infty \int_{-\infty}^\infty [u\phi_t + f(u)\phi_x + g(x - ct, u)\phi](x, t)dxdt + \int_{-\infty}^\infty u_0(x)\phi(x, 0)dx = 0.
$$

Several physical situations can be modeled as systems of hyperbolic conservation laws with a source such as a nozzle flow [7], [17], [19], [20], [21] and a moving magnetic field for magneto-hydrodynamics (MHD) [12]. As a prototype for systems of hyperbolic conservation laws with a source, we consider a quasi–one-dimensional nozzle flow model:

$$
\begin{aligned}
\frac{\partial \rho}{\partial t} + \frac{\partial(\rho u)}{\partial x} &= -\frac{A'(x)}{A(x)}(\rho u), \quad (x, t) \in R \times R_+, \\
\frac{\partial(\rho u)}{\partial t} + \frac{\partial(\rho u^2 + P)}{\partial x} &= -\frac{A'(x)}{A(x)}(\rho u^2),
\end{aligned}
$$

*Received by the editors May 24, 2000; accepted for publication (in revised form) February 28, 2001; published electronically July 19, 2001.
http://www.siam.org/journals/sima/33-2/37304.html

†Department of Mathematics, Serra Mall, Building 380, Stanford University, Stanford, CA, 94305 (syha@math.stanford.edu).

$$\frac{\partial(\rho E)}{\partial t} + \frac{\partial(\rho u E + P u)}{\partial x} = -\frac{A'(x)}{A(x)}(\rho u E + P u),$$
$$P = P(e, \rho),$$

where $A(x)$ is the cross sectional area of a nozzle, ρ the density, u the velocity, P the pressure, e the internal energy, and $E = e + \frac{u^2}{2}$ the total energy of a gas. The existence theory of a global weak solution for the system (1.1) was first established in [19] based on the modified Glimm scheme and the wave tracing method, and recently the global existence of weak solutions for L^∞ initial data was proved by using the compensated compactness in [7]. The basic idea for the construction of approximate solutions is to alternately use the Riemann solutions for a corresponding homogeneous system ($g(x, u) = 0$) and solutions traveling with a speed c. So far, most stability analyses for (1.1) have been carried out in bounded variation context [17], [19], [20], [21]. On the other hand, L^1 stability has been studied for scalar conservation laws [13], [14], [25].

Recently, there was a breakthrough for L^1 stability for systems of homogeneous conservation laws. So far, there are two different approaches. Bressan's approach is based on comparison and homotopy of two infinitesimally close solutions [2], [3], [4], [5]. In contrast, Liu and Yang's approach uses a robust nonlinear functional [23], [24]. In this paper, we adopt the latter approach which is based on the construction of a robust nonlinear functional $H[u, v]$ equivalent to the L^1 distance of u and v and nonincreasing in time. Without loss of generality, we may assume that the source has speed $c = 0$, i.e.,

$$u_t + f(u)_x = g(x, u).$$

As shown in [19], the stability of the Glimm solutions for (1.1) is subject to the following three essential mechanisms:

1. the genuine nonlinearity;
2. the nonresonance between hyperbolic waves and stationary waves;
3. the localization of the source in x, i.e., $\mathrm{supp}_x\{g(x, u)\}$ is compact.

Because of the above three mechanisms, hyperbolic waves are subsonic or supersonic; therefore, eventually, they will be away from the support of source. So they will be stabilized as hyperbolic waves for homogeneous systems. For the resonance case,[1] the interaction between hyperbolic waves and stationary waves will be quite complicated. In this case, the geometry of a nozzle is very important, as is shown for (1.1) with respect to special data [17], [20] and for a scalar model with respect to general data [21].

Based on the above three main mechanisms, we impose the following conditions on (1.1).

Main assumptions.

1. The system (1.1) is strictly hyperbolic.
   Let $\lambda_i(u), (i \in \{1, \dots, n\})$ be distinct real eigenvalues of $f'(u)$ and let $r_i(u)$ ($l_i(u)$) be corresponding right (left) eigenvectors of $f'(u)$, i.e.,

$$f'(u) r_i(u) = \lambda_i(u) r_i(u), \quad \lambda_1(u) < \cdots < \lambda_n(u),$$
$$l_i(u) f'(u) = \lambda_i(u) l_i(u), \quad l_i \cdot r_j = \delta_{ij}.$$

[1] One of the characteristic speeds is close to that of the source.

2. The source is not resonant with the conservation laws, that is, there exists $j_0 \in \{1, \dots, n-1\}$ such that

$$\lambda_1(u) < \cdots < \lambda_{j_0}(u) < 0 < \lambda_{j_0+1}(u) < \cdots < \lambda_n(u) \text{ for all } u \in \mathcal{N}.$$

3. Each characteristic field $(\lambda_j(u), r_j(u))$ is either genuinely nonlinear (g.n.l.) or linearly degenerate (l.d.g.) in the sense of [15]:

$$\begin{aligned}(\lambda_j(u), r_j(u)) \quad &\text{is g.n.l.} \iff \nabla\lambda_j(u) \cdot r_j(u) \neq 0 \quad \text{for all } u \in \mathcal{N},\\ (\lambda_j(u), r_j(u)) \quad &\text{is l.d.g.} \iff \nabla\lambda_j(u) \cdot r_j(u) \equiv 0.\end{aligned}$$

4. The total variation of initial data $u_0(x)$ is sufficiently small,

$$T.V_x(u_0(x)) \leq T.V << 1,$$

for a positive constant $T.V$ depending only on (1.1).

5. $g(x,p)$ is piecewise differentiable in x and continuously differentiable in p, and has compact support in x, and is sufficiently weak:

$$\begin{aligned}&g(x,p) \equiv 0, \quad x \notin [0,1] \ \text{ and } g(x,p) \neq 0, \quad x \in (0,1),\\ &G(x) \equiv \sup_{p\in\mathcal{N}} \left\{ |g(x,p)| + \left\| \frac{\partial g(x,p)}{\partial p} \right\| \right\}, \quad G_1 \equiv \|G(\cdot)\|_{L^1(R)},\\ &G_0 \equiv \|G(\cdot)\|_{L^\infty(R)}, \ G_0 + G_1 << 1.\end{aligned}$$

For (1.1) without a smallness assumption on the source, local L^1 stability in time was studied in [9], and when the source $g(x,u)$ depends only on u, under the dissipation condition on the source, L^1 stability was studied in [1]. The main theorem of this paper is as follows.

Theorem 1.2. *Let $u(x,t)$ and $v(x,t)$ be two weak solutions obtained by the Glimm scheme corresponding to initial data u_0 and v_0, respectively. Then under the main assumptions, we have*

$$||u(\cdot,t) - v(\cdot,t)||_{L^1(R)} \leq C||u_0(\cdot) - v_0(\cdot)||_{L^1(R)}, \quad t \geq 0,$$

where C is a generic constant which is independent of t.

The paper is organized as follows. In section 2, we review the basic theory of hyperbolic conservation laws and in section 3, we briefly discuss the simplified wave patterns [24] on which a nonlinear functional will be explicitly defined. In section 4, we study scalar conservation laws with a moving source. This section illustrates the necessity of a new functional which takes care of a source effect on the L^1 distance and treats an entropy functional. In section 5, we construct a nonlinear functional which contains a new functional measuring a source effect on the L^1distance and by using this functional, we prove L^1 stability for Glimm solutions.

2. Preliminaries. In this section, we review some of the basics for systems of hyperbolic conservation laws,

$$\begin{aligned}u_t + f(u)_x = 0, &\quad (x,t) \in R \times R_+,\\ u(x,0) = u_0(x), &\quad x \in R.\end{aligned} \tag{2.1}$$

The Riemann problem for (2.1) is the initial value problem with simple jump initial data

$$u(x,0) = \begin{cases} u_l, & x < 0,\\ u_r, & x > 0.\end{cases}$$

It is well known [15] that the Riemann solution is a function of $\frac{x}{t}$ and consists of $n+1$ intermediate constant states $\{u_l = u_0, u_1, u_2, \dots, u_n = u_r\}$ which are connected by shock waves, rarefaction waves, or contact discontinuities.

In the following, we define an ith rarefaction curve $R_i(u_0)$ and an ith shock curve $H_i(u_0)$.

$$\begin{aligned} R_i(u_0) &\equiv \text{ the integral curve of a vector field } r_i \cdot \nabla_u \text{ passing through } u_0, \\ H_i(u_0) &\equiv \{u \in R^n : \lambda_i(u_0, u)(u - u_0) = f(u) - f(u_0), \text{ for some scalar } \lambda_i(u_0, u)\}. \end{aligned}$$

For a l.d.g. characteristic field $(\lambda_i(u), r_i(u))$, it is well known [15] that $H_i(u_0) = R_i(u_0)$. We parameterize these curves by the arc length ξ and we divide the ith shock curve $H_i(u_0)$ and the ith rarefaction curve $R_i(u_0)$ as follows:

$$\begin{aligned} H_i^+(u_0) &\equiv \{u \in H_i(u_0) : \lambda_i(u) > \lambda(u_0, u) > \lambda_i(u_0)\}, \\ H_i^-(u_0) &\equiv \{u \in H_i(u_0) : \lambda_i(u) < \lambda(u_0, u) < \lambda_i(u_0)\}, \\ R_i^+(u_0) &\equiv \{u \in R_i(u_0) : \lambda_i(u) > \lambda_i(u_0)\}, \\ R_i^-(u_0) &\equiv \{u \in R_i(u_0) : \lambda_i(u) < \lambda_i(u_0)\}; \end{aligned}$$

moreover, we define an ith wave curve as follows:

$$W_i(u_0) \equiv \begin{cases} H_i^-(u_0) \cup R_i^+(u_0) & \text{if } i\text{th characteristic field is g.n.l.,} \\ H_i(u_0) = R_i(u_0) & \text{if } i\text{th characteristic field is l.d.g.} \end{cases}$$

Then by the second order contact of the ith shock curve and the ith rarefaction curve at u_0, the ith wave curve $W_i(u_0)$ is a C^2-curve [15].

THEOREM 2.1 (see [15]). *Suppose that* (2.1) *is strictly hyperbolic and each characteristic field is g.n.l. or l.d.g. If u_l and u_r are sufficiently close, then the Riemann problem for* (2.1) *has a unique solution in the class of elementary waves* $(u_{i-1}, u_i), u_i \in W_i(u_{i-1}), i = 1, 2, \dots, n, u_0 = u_l, u_n = u_r$.

For any i-wave $\alpha_i = (u_{i-1}, u_i)$, a signed strength $[\alpha_i]$ is defined as follows:

$$[\alpha_i] \equiv \mu_i(u_i) - \mu_i(u_{i-1}),$$

where μ_i is any nonsingular parameter along the ith wave curve $W_i(u_0)$ such that a shock wave has negative signed strength, whereas a rarefaction wave has positive signed strength. Approximate solutions for (1.1) are constructed by Riemann solutions for the corresponding homogeneous system and local steady solutions of (1.1) as building blocks [10], [18], [19]. For the details, we refer to [19].

DEFINITION 2.2. *Let $\{a_j\}_0^\infty$ be a sequence in* $(0, 1)$.

$$\{a_j\}_0^\infty \textit{ is equidistributed or (uniformly distributed)} \iff \lim_{N\to\infty} \frac{B(a_j, N, I)}{N} = |I|,$$

where $B(a_j, N, I) = |\{j : a_j \in I, 0 \le j \le N\}|$, for any subinterval I of $(0, 1)$ and $|I|$ denotes the length of the interval.

DEFINITION 2.3. *Let α_i and β_j be two i, j waves, respectively, such that α_i is located to the left of β_j.*

α_i and β_j are approaching if and only if $i > j$ or $i = j$ and at least one of them is a shock.

For later use, we state the approximate Rankine–Hugoniot condition for the approximate solutions.

Theorem 2.4 (see [19]). *Let (u_l, u_r) be a shock wave issued from (hr, ks) in the construction of approximate solutions and let $u_l(x,t)$ and $u_r(x,t)$ be two steady solutions with initial data u_l and u_r, respectively, along the $\frac{x-hr}{t-ks} = \lambda(u_l, u_r), ks < t < (k+1)s$. Then*

$$\lambda(u_l, u_r)(u_r(x,t) - u_l(x,t)) = f(u_r(x,t)) - f(u_l(x,t)) + \mathcal{O}(1)\left(\int_{hr}^{x} G(\xi)d\xi\right)|u_r - u_l|.$$

3. Known results. In this section, we study simplified wave patterns which will be used for the construction of a nonlinear functional $H(t)$ in section 5. These simplified wave patterns are a sort of generalization of those in [24], and we review the modified Glimm functional in [19].

3.1. Simplified wave patterns. Unlike linear waves, nonlinear waves change their strengths and speeds due to interactions and cancellations. The wave tracing is a book-keeping scheme of subdividing elementary waves in the approximate solution $u_r(x,t)$ so that the evolution of each subwave can be studied more definitely. After partitioning the waves in the approximate solutions [18], [19], we replace $u_r(x,t)$ by a simplified wave pattern $\bar{u}_r(x,t)$ consisting of nonlinear waves with deterministic speeds in each small time zone. Let us set

$$\begin{aligned}
&\text{time zone } \Lambda(t_1,t_2) \equiv \{(x,t) : -\infty < x < \infty, t_1 \le t < t_2\},\\
&\text{interaction measure } Q(t_1,t_2) \equiv \sum_{\Delta_{mn}\in\Lambda(t_1,t_2)} Q(\Delta_{mn}),\\
&\text{cancellation measure } C(t_1,t_2) \equiv \sum_{\Delta_{mn}\in\Lambda(t_1,t_2)} C(\Delta_{mn}),
\end{aligned}$$

where a local interaction measure $Q(\Delta_{mn})$ and a cancellation measure $C(\Delta_{mn})$ are defined as follows. Let Δ_{mn} be a diamond whose vertices are $((m-1)r + a_j r, ns)$, $(mr + a_j r, ns)$, $(mr, (n+\frac{1}{2})s)$, and $(mr, (n-\frac{1}{2})s)$.

$$\begin{aligned}
Q(\Delta_{mn}) &\equiv \sum_{(\alpha_i,\beta_j):app} \{|\alpha_i||\beta_j| : \ \alpha_i \text{ and } \beta_j \text{ pass through } \Delta_{mn}\},\\
C(\Delta_{mn}) &\equiv \sum_{(\alpha_i,\beta_j):app} \left\{\frac{|\alpha_i| + |\beta_i| - |\alpha_i + \beta_j|}{2} : \ \alpha_i \text{ and } \beta_i \text{ pass through } \Delta_{mn}\right\},
\end{aligned}$$

where (α_i, β_j) : *opp* denotes the approaching pair (α_i, β_j) defined in Definition 2.3. Without confusion, (u_{i-1}, u_i) denotes the i-wave or sometimes the difference of its end states. For the details and the motivation for the wave partition and tracing, we refer to [18], [19], [24]. The waves in $u_r(x,t), (x,t) \in \Lambda(t_1,t_2)$ consist of two different types of past history, i.e., the primitive waves issued from $t = t_1$ or waves which are generated by nonlinear interactions. Moreover, the waves also have two different futures (surviving in the future or cancelled in the future). Therefore, by refining waves in the approximate solutions in each small time zone, we can make waves into subwaves which will completely survive or be cancelled completely in the time zone $\Lambda(t_1,t_2)$ so that we can study the evolution of its subwaves definitely.

Let $\{a_j\}_0^\infty$ be an equidistributed sequence. Let ϵ be small and T be given; then let us set $N = \frac{1}{\epsilon}$, and let us choose M such that

$$(N-1)Ms < T \le NMs.$$

Without loss of generality, we may assume that N and M are integers; then it is easy to see that for a fixed N, as $s \to 0+$,

$$M \to \infty, \quad Ms \leq \frac{T\epsilon}{1-\epsilon}, \quad NMs \leq \frac{T}{1-\epsilon}.$$

We divide the interval $(0,1)$ into N equal subintervals with length ϵ. Let $\{I_i\}_1^{2^N}$ be the power set of such subintervals. Let us set

$$\delta = \sup_{1\leq p\leq N, 1\leq i\leq 2^N} \left\{ \frac{B(a_{m+(p-1)M}, M, I_i)}{M} - |I_i| \right\},$$

where $|I_i|$ denotes the Lebesgue measure of I_i. Then by the equidistributedness of $\{a_j\}$,

$$\lim_{M\to\infty} \delta = 0 \quad \text{ for any } \epsilon.$$

We partition the elementary waves in an approximate solution $u_r(x,t)$ into subwaves so that a rarefaction wave consists of rarefaction shocks whose maximal strength is ϵ [6], [18], [19], [23], [24]. In what follows, let us set

$$\Lambda_p \equiv \Lambda((p-1)Ms, pMs), \quad p = 1, \dots, N.$$

Based on the partitioned approximate solutions, we define a simplified wave pattern which consists of surviving nonlinear waves with fixed speeds in each small time zone Λ_p. This simplified wave pattern is a generalization of that in [24]. In [24], the simplified wave pattern consists of piecewise constant states, whereas our simplified wave pattern consists of piecewise stationary solutions. The construction of our simplified wave patterns is reviewed below.

Since $u_r(x,t)$ is of bounded variation, for a given small number ϵ, we can find E such that $T.V\{u_r(x,t) : |x| > E\} < \epsilon$. Next, we replace the $u_r(x,t)$ on $x < -E$ or $x > E$ by the values $\lim_{x\to-\infty} u(x,t)$ or $\lim_{x\to\infty} u(x,t)$, respectively. Hence we have a finite number of surviving waves on $[-E, E]$. Let us denote surviving i-waves by $v_i^1, \dots, v_i^N$. For each i-wave v_i^k, the location of it which is randomly determined by the sequence $\{a_j\}$ is now replaced by the line connecting its locations at time $t = (p-1)Ms+$ and $t = pMs-$. Let us denote its speed by $\lambda^*(v_i^k)$. As with the approximate solutions, the waves are connected to each other by stationary waves. Then it is noted that no i-waves do not cross each other in Λ_p. With regard to the secondary waves such as nonsurviving waves in $u_r(x,t)$ and generated waves from the nonlinear interactions in Λ_p, we do not keep track of them inside Λ_p. This generates an error in the L^1-norm which vanishes eventually, but in the beginning of the next time zone Λ_{p+1}, we consider those secondary waves. Therefore, the waves in the simplified wave pattern $\bar{u}_r(x,t)$ move in a deterministic way, but their end states evolve according to the stationary solution.

As a generalization of Theorem 5.3 in [24], we have the following theorem.

THEOREM 3.1 (see [24]). *There exists a simplified wave pattern $\bar{u}_r(x,t)$ consisting of a finite number of nonlinear waves $\{\bar{\alpha}\}$ in each time zone Λ_p and a large constant E such that the following hold.*

There exists a one-to-one correspondence $\alpha \to \bar{\alpha}$ between the surviving waves in $|x| < E$ and $K = \{\bar{\alpha}\}$ such that

1. $\sum_\alpha |\alpha - \bar{\alpha}| = \mathcal{O}(1)\{(Q_0 + Q_1 + C)(\Lambda_p) + \epsilon\}$,

2. $\sum_\alpha |\alpha||\lambda(\alpha) - \lambda^*(\bar{\alpha})| = \mathcal{O}(1)\{(Q_0 + Q_1 + C)(\Lambda_p) + \delta + \epsilon\}$,
3. $\sum\{|\alpha| : \alpha$ *is a secondary wave* $\} = \mathcal{O}(1)(Q_0 + Q_1 + C)(\Lambda_p)$,
4. $\bar{u}_r(x, (p-1)Ms) - u_r(x, (p-1)Ms) = 0$ *for* $|x| < E$,
5. $\int_{|x|>E} |\bar{u}_r(x, (p-1)Ms) - u_r(x, (p-1)Ms)|dx + \sum\{|\alpha| : \alpha \in u_r(x, (p-1)Ms), |x| > E\} < (T.V + G_1)\epsilon$.

3.2. The modified Glimm functional. Let $u_r(x,t)$ be a given approximate Glimm solution [19]. For such $u_r(x,t)$, we define the modified Glimm functional $F(u_r : t)$ as follows:

$$\begin{aligned} F(u_r : t) &= L(u_r : t) + KQ(u_r : t), \\ L(u_r : t) &= \sum\{|\alpha(t)| : \alpha \text{ is an elementary wave at time } t\}, \\ Q(u_r : t) &= Q_0(u_r : t) + Q_1(u_r : t), \\ Q_0(u_r : t) &= \sum\{|\alpha(t)||\beta(t)| : \alpha \text{ and } \beta \text{ are approaching }\}, \\ Q_1(u_r : t) &= \sum_{\lambda(\alpha(t))>0} \left\{ |\alpha(t)| \int_{hr}^{\infty} G(x)dx : \alpha \text{ is issued from } (hr, ks) \right\} \\ &\quad + \sum_{\lambda(\alpha(t))<0} \left\{ |\alpha(t)| \int_{-\infty}^{hr} G(x)dx : \alpha \text{ is issued from } (hr, ks) \right\}, \end{aligned}$$

where K is a large positive constant to be determined later.

Remark. The linear part $L(u_r : t)$ measures the strength of the waves at time t; therefore, $L(u_r : t)$ is equivalent to the total variation of the approximate solution at time t, and the quadratic part $Q(u_r : t)$ measures the potential interaction between waves. $Q_0(u_r : t)$ measures the potential interaction between hyperbolic waves, and $Q_1(u_r : t)$ measures the potential interaction between hyperbolic waves and stationary waves.

Then by the local interaction estimates of waves [19], we have the following decay estimates of the modified Glimm functional.

LEMMA 3.2 (see [19]). *Suppose that the total variation of initial data is sufficiently small and G_0 and G_1 are small enough. Then*

$$\begin{aligned} Q(u_r : \Lambda(0,t)) &\le 2(Q(u_r : 0) - Q(u_r : t)), \\ F(u_r : t+) - F(u_r : t-) &\le -\frac{1}{2}Q(u_r : t), \\ F(u_r : t) &\le F(u_r : 0) \quad \textit{for } t \ge 0. \end{aligned}$$

4. Scalar conservation laws with a source. In this section, in order to illustrate a functional which takes care of a source effect and an entropy functional in [22], we consider the scalar nozzle flow model which was introduced in [21]. We study time change of the L^1 distance between two solutions and a generalized entropy functional for scalar convex conservation laws with a source:

$$\begin{aligned} &u_t + f(u)_x = g(x,u), \quad (x,t) \in R \times R_+, \\ &f'(u) > 0,\ f''(u) > 0, \quad \mathrm{supp}_x(g(x,u)) = [0,1]. \end{aligned} \tag{4.1}$$

Unlike homogeneous scalar conservation laws, because of the source, the L^1 distance between two weak solutions may not be a contraction. Under the above assumptions of the source, we verify that in general, the L^1 distance is not a contraction. Moreover, we show that the $H(t) = ||u(\cdot,t) - v(\cdot,t)||_{L^1(R)}$ for the homogeneous case is not applicable to a nonhomogeneous case (Theorem 4.3). Hereafter, we will use the following notations:

$u_r(x,t)$: the approximate solution by the modified Glimm scheme
with a space mesh size r,
$\bar{u}_r(x,t)$: the simplified wave pattern corresponding to $u_r(x,t)$,
$\alpha_j^i(t)$: the jth i-wave at time t, $x(\alpha(t))$: the location of the wave $\alpha(t)$,
$\lambda(\alpha(t))$: the exact speed of $\alpha(t)$, $\dot{x}(\alpha(t))$: the approximate speed of $\alpha(t)$,
$J(\bar{u}_r)$: the set of all waves in $\bar{u}_r$, $J(\bar{v}_r)$: the set of all waves in $\bar{v}_r$,
$J \equiv J(\bar{u}_r) \cup J(\bar{v}_r)$, $q^{\pm}(\alpha) \equiv v(x(\alpha)\pm,t) - u(x(\alpha)\pm,t)$,
$\lambda(q^{\pm}(\alpha)) \equiv$ the speed of $(u(x(\alpha)\pm,t), v(x(\alpha)\pm,t))$.

We will also use the abbreviated notations $\bar{u}_r(x)$ and $\bar{v}_r(x)$ instead of $\bar{u}_r(x,t)$ and $\bar{v}_r(x,t)$, respectively. Let $u(x,t)$ and $v(x,t)$ be two solutions constructed by the modified Glimm scheme [19] corresponding to initial data $u_0(x)$ and $v_0(x)$, respectively, such that

$$\lim_{r\to 0} u_r(x,t) = u(x,t), \ \lim_{r\to 0} v_r(x,t) = v(x,t) \ \text{ in } L^1_{loc}(R \times R_+),$$
$$\text{and } \|u_0(x) - v_0(x)\|_{L^1(R)} < \infty.$$

For a given noninteraction time t, let us denote the set of all waves by $\{\alpha_i\}_1^m$ such that

$$-\infty < x(\alpha_1) < x(\alpha_2) < \cdots < x(\alpha_m) < \infty.$$

Then it is easy to see that

$$\text{(4.2a)} \qquad \frac{d}{dt}|\bar{u}_r(x) - \bar{v}_r(x)| = 0 \text{ on } (x(\alpha_i), x(\alpha_{i+1})), \ i = 1, \ldots, m-1,$$
$$\text{(4.2b)} \qquad \bar{u}_r(x) = \bar{v}_r(x) \text{ on } (-\infty, x(\alpha_1)) \ \text{ and } (x(\alpha_m), \infty).$$

On the other hand, we have

$$\begin{aligned}\text{(4.3)} \ \ \frac{d}{dt}\int_{x(\alpha_i)}^{x(\alpha_{i+1})} |\bar{u}_r(x) - \bar{v}_r(x)|dx &= \int_{x(\alpha_i)}^{x(\alpha_{i+1})} \frac{d}{dt}|\bar{u}_r(x) - \bar{v}_r(x)|dx \\ &+ \dot{x}(\alpha_{i+1})|\bar{u}_r(x(\alpha_{i+1})-) - \bar{v}_r(x(\alpha_{i+1})-)| - \dot{x}(\alpha_i)|\bar{u}_r(x(\alpha_i)+) - \bar{v}_r(x(\alpha_i)+)| \\ &= \dot{x}(\alpha_{i+1})|q^-(\alpha_{i+1})| - \dot{x}(\alpha_i)|q^+(\alpha_i)| \ \text{ by } (4.2a).\end{aligned}$$

It follows from (4.2b) and (4.4) that

$$
\begin{aligned}
&\frac{d}{dt}\|\bar{u}_r(\cdot,t)-\bar{v}_r(\cdot,t)\|_{L^1(R)} = \sum_{i=1}^{m-1}\frac{d}{dt}\int_{x(\alpha_i)}^{x(\alpha_{i+1})}|\bar{u}_r(x)-\bar{v}_r(x)|dx\\
&=\sum_{i=1}^{m}\{\dot{x}(\alpha_i)(|q^-(\alpha_i)|-|q^+(\alpha_i)|)\}\\
&=\sum_{i=1}^{m}\lambda(\alpha_i)(|q^-(\alpha_i)|-|q^+(\alpha_i)|)+\mathcal{O}(1)MsG_0\sum_{i=1}^{m}|\alpha_i|\\
&=\sum_{i=1}^{m-1}\{\lambda(\alpha_{i+1})|q^-(\alpha_{i+1})|-\lambda(\alpha_i)|q^+(\alpha_i)|\}+\mathcal{O}(1)MsG_0\sum_{1}^{m}|\alpha_i|\\
&=\sum_{i=1}^{m-1}\{(\lambda(\alpha_{i+1})-\lambda(q^-(\alpha_{i+1})))|q^-(\alpha_{i+1})|-(\lambda(\alpha_i)-\lambda(q^+(\alpha_i)))|q^+(\alpha_i)|\}\\
&+\sum_{i=1}^{m-1}\{\lambda(q^-(\alpha_{i+1}))|q^-(\alpha_{i+1})|-\lambda(q^+(\alpha_i))|q^+(\alpha_i)|\}+\mathcal{O}(1)MsG_0\sum_{1}^{m}|\alpha_i|\\
&\equiv I+II+\mathcal{O}(1)MsG_0\sum_{1}^{m}|\alpha_i|,
\end{aligned}
$$

where

$$
\begin{aligned}
I &\equiv \sum_{i=1}^{m-1}\{(\lambda(\alpha_{i+1})-\lambda(q^-(\alpha_{i+1})))|q^-(\alpha_{i+1})|-(\lambda(\alpha_i)-\lambda(q^+(\alpha_i)))|q^+(\alpha_i)|\},\\
II &\equiv \sum_{i=1}^{m-1}\{\lambda(q^-(\alpha_{i+1}))|q^-(\alpha_{i+1})|-\lambda(q^+(\alpha_i))|q^+(\alpha_i)|\},
\end{aligned}
$$

and we have used the fact that $\dot{x}(\alpha_i)-\lambda(\alpha_i)=\mathcal{O}(1)MsG_0$ and $|q^-(\alpha_i)|-|q^+(\alpha_i)|\leq|\alpha_i|$. Let us recall that for scalar convex conservation laws, unlike for systems, nonlinear interactions do not generate new waves. For simplicity, we assume that the initial data is constant outside a bounded interval. Then at time $t=0$, since there are only a finite number of waves in $\bar{u}_r(x,0)$ and $\bar{v}_r(x,0)$ and each wave has a positive speed, there exists a finite time T_{es} such that

$$x(\alpha_i(t))>1 \quad \text{for all } i=1,\dots,m \text{ and } t\geq T_{es}.$$

In the following two lemmas, we estimate I and II separately.

LEMMA 4.1. *The quantity I satisfies the following estimate:*

$$I\leq -C_1\sum\{|q^-(\alpha_i)||q^+(\alpha_i)|\}+|\mathcal{O}(1)|\epsilon(T.V+G_1),$$

where $|\mathcal{O}(1)|$ and C_1 depend only on (4.1), ϵ is the upper bound of the strength of rarefaction shocks in the simplified wave pattern, and the summation is over all waves in J which cross the other solution.

Proof. Since I is a finite sum, by rearrangement, we can rewrite I as follows:

$$\begin{aligned} I &= \sum_{i=1}^{m}\{(\lambda(\alpha_i)-\lambda(q^-(\alpha_i)))|q^-(\alpha_i)| - (\lambda(\alpha_i)-\lambda(q^+(\alpha_i)))|q^+(\alpha_i)|\} \\ &= \sum_{i=1}^{m} A(\alpha_i), \end{aligned}$$

where $A(\alpha_i) = (\lambda(\alpha_i)-\lambda(q^-(\alpha_i)))|q^-(\alpha_i)| - (\lambda(\alpha_i)-\lambda(q^+(\alpha_i)))|q^+(\alpha_i)|$.

We consider a generic case (the locations of discontinuities in $\bar{v}_r$ and $\bar{u}_r$ are not coincident).

Case 1. $\alpha_i = (v_-, v_+) \in J(\bar{v}_r)$, and assume that $\bar{u}_r$ is continuous at $x = x(\alpha_i)$. We claim that

$$A(\alpha_i) = \begin{cases} 0, & q^-(\alpha_i)q^+(\alpha_i) \geq 0, \\ -|\mathcal{O}(1)||q^+(\alpha_i)||q^-(\alpha_i)|, & q^-(\alpha_i) \geq 0,\ q^+(\alpha_i) \leq 0, \\ |\mathcal{O}(1)|\epsilon|\alpha_i|, & q^-(\alpha_i) \leq 0, q^+(\alpha_i) \geq 0. \end{cases} \tag{4.4}$$

By the Rankine–Hugoniot condition and definition of $q^{\pm}(\alpha_i)$,

$$\lambda(\alpha_i)[\alpha_i] + \lambda(q^-(\alpha_i))q^-(\alpha_i) = \lambda(q^+(\alpha_i))q^+(\alpha_i). \tag{4.5}$$

Subcase 1.1. $q^-(\alpha_i) \geq 0,\ q^+(\alpha_i) \geq 0$. We have

$$\begin{aligned} A(\alpha_i) &= (\lambda(\alpha_i)-\lambda(q^-(\alpha_i)))|q^-(\alpha_i)| - (\lambda(\alpha_i)-\lambda(q^+(\alpha_i)))|q^+(\alpha_i)| \\ &= \lambda(\alpha_i)(q^-(\alpha_i)-q^+(\alpha_i)) - \lambda(q^-(\alpha_i))q^-(\alpha_i) + \lambda(q^+(\alpha_i))q^+(\alpha_i) \\ &= -\lambda(\alpha_i)[\alpha_i] - \lambda(q^-(\alpha_i))q^-(\alpha_i) + \lambda(q^+(\alpha_i))q^+(\alpha_i) = 0 \text{ by (4.5)}. \end{aligned}$$

Subcase 1.2. $q^-(\alpha_i) \leq 0,\ q^+(\alpha_i) \leq 0$. By the same analysis as Subcase 1.1, we have $A(\alpha_i) = 0$.

Subcase 1.3. $q^-(\alpha_i) \geq 0, q^+(\alpha_i) \leq 0$. By the convexity of the flux function, we have

$$\begin{aligned} &\lambda(\alpha_i) - \lambda(q^-(\alpha_i)) = -|\mathcal{O}(1)||q^+(\alpha_i)|, \\ &\lambda(\alpha_i) - \lambda(q^+(\alpha_i)) = |\mathcal{O}(1)||q^-(\alpha_i)|, \\ &A(\alpha_i) = (\lambda(\alpha_i)-\lambda(q^-(\alpha_i)))|q^-(\alpha_i)| - (\lambda(\alpha_i)-\lambda(q^+(\alpha_i)))|q^+(\alpha_i)| \\ &\qquad = -|\mathcal{O}(1)||q^+(\alpha_i)||q^-(\alpha_i)|. \end{aligned}$$

Subcase 1.4. $q^-(\alpha_i) \leq 0,\ q^+(\alpha_i) \geq 0$. In this case, since the upper bound of a rarefaction shock is ϵ, we have

$$\max\{|q^+(\alpha_i)|, |q^-(\alpha_i)|\} \leq |\alpha_i| < \epsilon.$$

Again, by the convexity of the flux function, we have

$$\begin{aligned} &\lambda(\alpha_i) - \lambda(q^-(\alpha_i)) = |\mathcal{O}(1)||q^+(\alpha_i)|, \\ &\lambda(\alpha_i) - \lambda(q^+(\alpha_i)) = -|\mathcal{O}(1)||q^-(\alpha_i)|, \\ &A(\alpha_i) \leq |\mathcal{O}(1)||q^+(\alpha_i)||q^-(\alpha_i)| \leq |\mathcal{O}(1)|\epsilon|\alpha_i|. \end{aligned}$$

Case 2. $\alpha_i = (u_-, u_+) \in J(\bar{u}_r)$, and $\bar{v}_r$ is continuous at $x = x(\alpha_i)$. We claim that

$$A(\alpha_i) = \begin{cases} 0, & q^-(\alpha_i)q^+(\alpha_i) \geq 0, \\ |\mathcal{O}(1)|\epsilon|\alpha_i|, & q^-(\alpha_i) \geq 0,\ q^+(\alpha_i) \leq 0, \\ -|\mathcal{O}(1)||q^+(\alpha_i)||q^-(\alpha_i)|, & q^-(\alpha_i) \leq 0, q^+(\alpha_i) \geq 0. \end{cases} \tag{4.6}$$

By the Rankine–Hugoniot condition and definition of $q^{\pm}(\alpha_i)$,

$$\lambda(\alpha_i)[\alpha_i] + \lambda(q^+(\alpha_i))q^+(\alpha_i) = \lambda(q^-(\alpha_i))q^-(\alpha_i). \tag{4.7}$$

By the same analysis using (4.7) as in Case 1, we get (4.6). From (4.3) and (4.6), we conclude that

$$I \leq -C_1 \sum \{|q^-(\alpha_i)||q^+(\alpha_i)|\} + |\mathcal{O}(1)|\epsilon(T.V + G_1),$$

where the summation is over all waves in J which cross the other solution. □

Lemma 4.2. *The quantity II satisfies the following estimate:*

$$II \leq \chi_{[0,T_{es}]}(t) \int_0^1 G(x)|\bar{u}_r(x) - \bar{v}_r(x)|dx,$$

where $\chi_{[0,T_{es}]}(t)$ is the characteristic function of the interval $[0, T_{es}]$.

Proof. By the construction of a simplified wave pattern, if $q(x_0, t) = 0$ for some $x_0 \in ((x(\alpha_i), x(\alpha_{i+1}))$, then $q(x, t) = 0$ on $(x(\alpha_i), x(\alpha_{i+1}))$. Therefore, we need to consider only two cases:

$$\text{either } q^+(\alpha_i) \geq 0,\ q^-(\alpha_{i+1}) \geq 0,\ \text{ or } q^+(\alpha_i) \leq 0,\ q^-(\alpha_{i+1}) \leq 0.$$

Let us set $II(\alpha_i, \alpha_{i+1}) \equiv \lambda(q^-(\alpha_{i+1}))|q^-(\alpha_{i+1})| - \lambda(q^+(\alpha_i))|q^+(\alpha_i)|$.

Case 1. $q^+(\alpha_i) \geq 0, q^-(\alpha_{i+1}) \geq 0$.

$$\begin{aligned} II(\alpha_i, \alpha_{i+1}) &= \lambda(q^-(\alpha_{i+1}))q^-(\alpha_{i+1}) - \lambda(q^+(\alpha_i))q^+(\alpha_i) \\ &= f(\bar{v}_r(x(\alpha_{i+1})-)) - f(\bar{u}_r(x(\alpha_{i+1})-)) \\ &\quad - \{f(\bar{v}_r(x(\alpha_i)-)) - f(\bar{u}_r(x(\alpha_i)))\} \\ &= \int_{x(\alpha_i)}^{x(\alpha_{i+1})} \{f(\bar{v}_r(x))_x - f(\bar{u}_r(x))_x\}dx \\ &= \int_{x(\alpha_i)}^{x(\alpha_{i+1})} \{g(x, \bar{v}_r(x)) - g(x, \bar{u}_r(x))\}dx. \end{aligned} \tag{4.8}$$

Case 2. $q^+(\alpha_i) \leq 0, q^-(\alpha_{i+1}) \leq 0$. By the same calculation as Case 1, we have

$$II(\alpha_i, \alpha_{i+1}) = \int_{x(\alpha_i)}^{x(\alpha_{i+1})} \{g(x, \bar{u}_r(x)) - g(x, \bar{v}_r(x))\}dx. \tag{4.9}$$

For $t \geq T_{es}$, since $g(x, \bar{v}_r(x)) = g(x, \bar{u}_r(x)) = 0$ on $(x(\alpha_i(t)), x(\alpha_{i+1}(t)))$, $II(\alpha_i, \alpha_{i+1}) = 0$. For $t < T_{es}$, $|g(x, \bar{v}_r(x)) - g(x, \bar{u}_r(x))| \leq G(x)|\bar{u}_r(x) - \bar{v}_r(x)|$ on $(x(\alpha_i), x(\alpha_{i+1}))$. Thus, in (4.8) and (4.9) we have

$$II \leq \chi_{[0,T_{es}]}(t) \int_0^1 G(x)|\bar{u}_r(x) - \bar{v}_r(x)|dx.$$

This completes the proof. □

By combining Lemmas 4.1 and 4.2, we have the following estimates on the L^1 distance between two simplified wave patterns.

Theorem 4.3. *There exists a positive constant C_1 depending on* (4.1) *such that*

$$\frac{d}{dt}\|\bar{u}_r(\cdot, t) - \bar{v}_r(\cdot, t)\|_{L^1(R)} \leq -C_1 \sum \{|q^-(\alpha_i)||q^+(\alpha_i)|\} + |\mathcal{O}(1)|\epsilon(T.V + G_1)$$

$$+ \chi_{[0,T_{es}]}(t) \int_0^1 G(x)|\bar{u}_r(x) - \bar{v}_r(x)|dx$$
$$+ \mathcal{O}(1)MsG_0 \sum_1^m |\alpha_i|,$$

where the summation is over all genuine shock waves in one of the solutions which cross the other solution.

Remark 4.1. 1. For the two exact solutions $u(x,t)$ and $v(x,t)$, we have

$$\frac{d}{dt}\|u(\cdot,t) - v(\cdot,t)\|_{L^1(R)} \leq -C_1 \sum\{|q^-(\alpha_i)||q^+(\alpha_i)|\} + \chi_{[0,T_{es}]}(t) \int_0^1 G(x)|u(x) - v(x)|dx.$$

2. $\chi_{[0,T_{es}]}(t) \int_0^1 G(x)|\bar{u}_r(x) - \bar{v}_r(x)|dx$ implies the source effect on the L^1 distance.

In the following lemma, we estimate the time-variation of a shock strength.

LEMMA 4.4. *Let $\alpha(t) = (u_-(t), u_+(t))$, $(p-1)Ms \leq t < pMs$ be a discontinuity in $\bar{u}_r(x,t)$. Then*

$$\frac{d|\alpha(t)|}{dt} = \mathcal{O}(1)G(x(\alpha(t)))|\alpha(t)|,$$

where $\mathcal{O}(1)$ depends only on the system (4.1).

Proof. Let $\gamma(t) = (x(t), t)$ be the locus of $\alpha(t)$ in $x - t$ plane. Then the curve $\gamma(t)$ is differentiable at a.e. $t \in ((p-1)Ms, pM)$. Since the shock has a constant speed and strength in the region $[0,1]^c$ by the construction, we consider only the case $x(t) \in [0,1]$. Let us set

$$u_-(t) \equiv u(x(t)-, t), \quad u_+(t) \equiv u(x(t)+, t).$$

For definiteness, assume that $u_-((p-1)Ms) > u_+((p-1)Ms)$; then $u_-(t) > u_+(t), t \in ((p-1)Ms, pMs)$ and $|\alpha(t)| = u_-(t) - u_+(t)$. Since $u_-(t)$ and $u_+(t)$ are local steady solutions of (4.1), we have

$$\frac{du_-(t)}{dt} = \frac{\partial u_-(t)}{\partial x}\dot{x}(t) = \frac{\dot{x}(t)}{f'(u_-(t))}g(x(t), u_-(t)), \tag{4.10a}$$

$$\frac{du_+(t)}{dt} = \frac{\partial u_+(t)}{\partial x}\dot{x}(t) = \frac{\dot{x}(t)}{f'(u_+(t))}g(x(t), u_+(t)). \tag{4.10b}$$

It follows from (4.10a)–(4.10b) that

$$\begin{aligned}\frac{d|\alpha(t)|}{dt} &= \frac{du_-(t)}{dx} - \frac{du_+(t)}{dx} = \frac{\dot{x}(t)}{f'(u_-(t))}g(x(t), u_-(t)) - \frac{\dot{x}(t)}{f'(u_+(t))}g(x(t), u_+(t)) \\ &= \dot{x}(t)\frac{\partial}{\partial p}\left(\frac{g(x,p)}{f'(p)}\right)\Bigg|_{(x(t),\theta(t))}(u_-(t) - u_+(t)) = \mathcal{O}(1)G(x(\alpha))|\alpha(t)|,\end{aligned}$$

where $\theta(t)$ is between $u_-(t)$ and $u_+(t)$, and we have used the fact that $\frac{\partial}{\partial p}(\frac{g(x,p)}{f'(p)}) = \mathcal{O}(1)G(x(\alpha)), \dot{x}(t) = \mathcal{O}(1)$. This completes the proof. ∎

Next we define an entropy functional $E[\bar{u}_r, \bar{v}_r]$ to obtain the third order decay estimate for two simplified wave patterns $\bar{u}_r$ and $\bar{v}_r$. This entropy functional will be used to capture the nonlinearity of a characteristic field. From now on, without confusion we rewrite $\bar{u}_r, \bar{v}_r$ as u, v, respectively, and define

$$(u-v)_+ \equiv \max\{u-v, 0\}, \qquad (u-v)_- \equiv \max\{-(u-v), 0\}.$$

Let us set

$$\begin{aligned} E(\alpha) &\equiv |\alpha| \cdot \begin{cases} \int_{x(\alpha)}^{\infty}(u-v)_+(x,t)dx + \int_{-\infty}^{x(\alpha)}(u-v)_-(x,t)dx, & \alpha \in J(u), \\ \int_{x(\alpha)}^{\infty}(v-u)_+(x,t)dx + \int_{-\infty}^{x(\alpha)}(v-u)_-(x,t)dx, & \alpha \in J(v), \end{cases} \\ E(t) &\equiv E[u(\cdot,t), v(\cdot,t)] = \sum_{\alpha \in J} E(\alpha). \end{aligned}$$

In the following theorem, we study the time-variation of the entropy functional.

Theorem 4.5. *Let $u(x,t)$ and $v(x,t)$ be two simplified wave patterns of* (4.1) *whose total variations are bounded by $\mathcal{O}(1)(T.V + G_1)$. Then the entropy functional $E(t)$ satisfies*

$$\begin{aligned} &\frac{d}{dt}E(t) \le -C_2 \sum_{\alpha \in J} |\alpha| \max\{q^+(\alpha)q^-(\alpha), 0\} + \mathcal{O}(1) \sum_{\alpha \in J} G(x(\alpha))E(\alpha) \\ &+ \chi_{[0,T_{es}]}(t) \sum_{\alpha \in J(v)} |\alpha| \left\{ \int_{x(\alpha)}^{\infty} G(x)(v-u)_+(x,t)dx + \int_{-\infty}^{x(\alpha)} G(x)(v-u)_-(x,t)dx \right\} \\ &+ \chi_{[0,T_{es}]}(t) \sum_{\alpha \in J(u)} |\alpha| \left\{ \int_{x(\alpha)}^{\infty} G(x)(u-v)_+(x,t)dx + \int_{-\infty}^{x(\alpha)} G(x)(u-v)_-(x,t)dx \right\} \\ &+ \mathcal{O}(1)(T.V + G_1)^2\epsilon + \mathcal{O}(1)MsG_0 \sum_1^m |\alpha_i|, \quad \text{a.e. } t \in ((p-1)Ms, pMs), \end{aligned}$$

where C_2 is a positive constant depending only on (4.1).

Proof. Let us assume that $\alpha \in J(v)$, and u is continuous at $x = x(\alpha)$. Then by the definition of $E(\alpha)$,

$$\begin{aligned} \frac{dE(\alpha)}{dt} &= \frac{d}{dt}\left[|\alpha| \left\{ \int_{-\infty}^{x(\alpha)} (v-u)_-(x,t)dx + \int_{x(\alpha)}^{\infty} (v-u)_+(x,t)dx \right\} \right] \\ &= |\alpha| \left\{ \frac{d}{dt}\int_{-\infty}^{x(\alpha)} (v-u)_-(x,t)dx + \frac{d}{dt}\int_{x(\alpha)}^{\infty} (v-u)_+(x,t)dx \right\} \\ &+ \frac{d|\alpha|}{dt} \left\{ \int_{-\infty}^{x(\alpha)} (v-u)_-(x,t)dx + \int_{x(\alpha)}^{\infty} (v-u)_+(x,t)dx \right\} \\ &\equiv I(\alpha) + II(\alpha). \end{aligned}$$

From Lemma 4.4, we have

$$II(\alpha) = \mathcal{O}(1)G(x(\alpha))E(\alpha). \tag{4.11}$$

Let $\{x_j(t)\}$ be the partition of R such that

$$\begin{aligned} v(x,t) > u(x,t), &\quad x_{2i}(t) < x < x_{2i+1}(t), \\ v(x,t) < u(x,t), &\quad x_{2i-1}(t) < x < x_{2i}(t). \end{aligned}$$

Then by the construction of simplified wave patterns, either $u(x,t)$ or $v(x,t)$ has a discontinuity at $x = x_j(t)$. In the following, $q^{\pm}(x_j(t))$ denotes a wave $(u(x_j(t)\pm, t)$, $v(x_j(t)\pm, t))$ or a difference $v(x_j(t)\pm, t) - u(x_j(t)\pm, t)$.

By the same analysis as in [24], we get the following estimate:

$$\begin{aligned}
&\frac{d}{dt}\sum_{\alpha\in J(v)} E(\alpha) \le -C_2 \sum_{\alpha\in J(v)} |\alpha| \max\{q^+(\alpha)q^-(\alpha),0\} + \mathcal{O}(1)(T.V + G_1)^2\epsilon \\
&+ \chi_{[0,T_{es}]}(t) \sum_{\alpha\in J(v)} |\alpha| \left\{ \int_{x(\alpha)}^{\infty} G(x)(v-u)_+(x,t)dx + \int_{-\infty}^{x(\alpha)} G(x)(v-u)_-(x,t)dx \right\} \\
&+ \mathcal{O}(1)MsG_0 \sum_1^m |\alpha_i| + \mathcal{O}(1) \sum_{\alpha\in J(v)} G(x(\alpha))E_2(\alpha).
\end{aligned}$$

Similar estimates hold for $\sum_{\alpha\in J(u)} E(\alpha)$. Hence, we have

$$\begin{aligned}
\frac{dE(t)}{dt} \le\ & -C_2 \sum_{\alpha\in J} |\alpha| \max\{q^+(\alpha)q^-(\alpha),0\} + \mathcal{O}(1)(T.V + G_1)^2\epsilon \\
&+ \chi_{[0,T_{es}]}(t) \sum_{\alpha\in J(v)} |\alpha| \left\{ \int_{x(\alpha)}^{\infty} G(x)(v-u)_+(x,t)dx + \int_{-\infty}^{x(\alpha)} G(x)(v-u)_-(x,t)dx \right\} \\
&+ \chi_{[0,T_{es}]}(t) \sum_{\alpha\in J(u)} |\alpha| \left\{ \int_{x(\alpha)}^{\infty} G(x)(u-v)_+(x,t)dx + \int_{-\infty}^{x(\alpha)} G(x)(u-v)_-(x,t)dx \right\} \\
&+ \mathcal{O}(1)MsG_0 \sum_1^m |\alpha_i| + \mathcal{O}(1) \sum_{\alpha\in J(v)} G(x(\alpha))E_2(\alpha).
\end{aligned}$$

This completes the proof. □

5. Systems of hyperbolic conservation laws with a moving source.

In this section, we construct a nonlinear functional which is equivalent to the L^1 distance between two Glimm solutions and nonincreasing in time. By using this nonlinear functional, we establish the L^1 stability of the Glimm solutions.

5.1. A nonlinear functional. We define a nonlinear functional $H(t)$ which is equivalent to the L^1 distance and nonincreasing in time. Our analysis makes use of the strict hyperbolicity and the genuine nonlinearity, the fact that the source has a compact support in x, and the nonresonance condition of the system (1.1). Let us set

$$\begin{aligned}
&\mathcal{G} \equiv \{u(x,t) \in BV(R\times R_+) : u \ \text{ is a solution corresponding to initial data } \ u_0\}, \\
&\mathcal{D} \equiv \{(u,v) \in \mathcal{G}\times\mathcal{G} : u_0 - v_0 \in L^1(R)\}.
\end{aligned}$$

Let $u(x,t)$ and $v(x,t)$ be two weak solutions of (1.1) such that

$$\lim_{r\to 0} u_r(x,t) = u(x,t), \quad \lim_{r\to 0} v_r(x,t) = v(x,t) \quad \text{in } L^1_{loc}(R\times R_+).$$

Let $\bar{u}_r(x,t)$ and $\bar{v}_r(x,t)$ be the simplified wave patterns corresponding to $u_r(x,t)$ and $v_r(x,t)$, respectively. For the time being, we will fix r and without confusion, we rewrite $\bar{u}_r(x,t)$ and $\bar{v}_r(x,t)$ as $u(x,t)$ and $v(x,t)$. For given $(x,t)\in R\times R_+$, we solve the Riemann problem for the corresponding homogeneous conservation laws of (1.1) with initial data $(u(x,t),v(x,t))$ by shock waves or rarefaction shocks, i.e.,

$$\omega_0(x,t) = u(x,t), \ \ \omega_n(x,t) = v(x,t), \ \ \omega_i(x,t) \in H_i(\omega_{i-1}(x,t)), \ i = 0,1,\dots,n.$$

Moreover, if necessary, by a linear transformation of $u = (u^1,\dots,u^n)$, we may assume that each coordinate function u^i is strictly increasing along the ith wave curve

$W_i(u)$. Let us define $q_i(x,t)$, the strength of ith wave $(\omega_{i-1}(x,t), \omega_i(x,t))$, by the ith component of $\omega_i(x,t) - \omega_{i-1}(x,t)$, i.e.,

$$q_i(x,t) \equiv (\omega_i(x,t) - \omega_{i-1}(x,t))^i, \quad i = 1, \ldots, n.$$

Since $\mathcal{N} \subset$ compact subset of R^n, it is easy to see that

$$\frac{1}{C_3}|u(x,t) - v(x,t)| \leq \sum_1^n |q_i(x,t)| \leq C_3|u(x,t) - v(x,t)|,$$

where C_3 is a large positive constant which is independent of t.

Let us consider i-wave $\alpha^i \in J$, and define the location of the i-wave α^i and the waves generated by the difference of u and v at both sides of the wave as follows:

$$\begin{aligned} x(\alpha^i) \equiv & \quad \text{the location of } i\text{-wave } \alpha^i, \\ q_j^{\pm}(\alpha^i) \equiv q_j(x(\alpha^i)&\pm, t), \quad \lambda_j^{\pm}(\alpha^i) \equiv \lambda_j(\omega_{j-1}(x(\alpha^i)\pm, t), \omega_j(x(\alpha^i)\pm, t)). \end{aligned}$$

For $j = i$, we use abbreviated notations $q^{\pm}(\alpha^i), \lambda^{\pm}(\alpha^i)$. The nonlinear functional $H(t)$ is the weighted linear combination of four component functionals: $L(t)$ measuring the L^1 distance between two solutions $u(x,t)$ and $v(x,t)$, $Q_d(t)$ measuring nonlinear couplings between waves of different characteristic families, $E(t)$ capturing the nonlinearity of the characteristic field due to the bifurcation of a shock curve and a rarefaction curve, and $Q_{so}(t)$ measuring the source effect on the L^1 distance.

In contrast with the Liu–Yang functional [24], our modified nonlinear functional is defined to capture the effect of the source on the L^1 distance. For this, we need to consider the potential interaction between the imaginary wave [2] $q_i(x,t)$ and stationary waves. Recall that the strength of the stationary waves is measured by the function $G(x)$. In order to fix the idea, let us consider an imaginary wave $q_j(x_0,t), j > j_0$. Since this wave has a positive speed, it will propagate to ∞; in doing so, it will interact with the stationary waves lying $x \geq x_0$. The same argument holds for $q_j(x_0,t), j \leq j_0$. So potential interactions between imaginary waves located at $x = x_0$ and stationary waves are

$$\sum_{j \leq j_0} |q_j(x_0,t)| \int_{-\infty}^{x(q_j)} G(\xi) d\xi + \sum_{j \geq j_0+1} |q_j(x_0,t)| \int_{x(q_j)}^{\infty} G(\xi) d\xi.$$

Based on this observation, we define a nonlinear functional which contains a new component functional $Q_{so}(t)$ in the following. First we define a nonlinear functional for two simplified wave patterns $u(x,t)$ and $v(x,t)$.

$$L(t) \equiv \sum_{j=1}^n L_j(t) \equiv \sum_{j=1}^n \int_{-\infty}^{\infty} |q_j(x,t)| dx,$$

$$Q_d(t) \equiv \sum_{\alpha^i \in J} Q_d(\alpha^i(t)) \equiv \sum_{\alpha^i \in J} |\alpha^i(t)| \left\{ \sum_{j>i} \int_{-\infty}^{x(\alpha^i)} |q_j(x,t)| dx + \sum_{j<i} \int_{x(\alpha^i)}^{\infty} |q_j(x,t)| dx \right\},$$

$$E(t) \equiv \sum_{\alpha^i \in J} E(\alpha^i(t)) \equiv \sum_{\alpha^i \in J} |\alpha^i(t)| \cdot \begin{cases} \int_{-\infty}^{x(\alpha^i)} q_i(x,t)_+ dx + \int_{x(\alpha^i)}^{\infty} q_i(x,t)_- dx, & \alpha^i \in J(u), \\ \int_{x(\alpha^i)}^{\infty} q_i(x,t)_+ dx + \int_{-\infty}^{x(\alpha^i)} q_i(x,t)_- dx, & \alpha^i \in J(v), \end{cases}$$

[2] These waves are not the real hyperbolic waves in weak solutions but virtual waves generated by the difference between two weak solutions.

$$
\begin{aligned}
Q_{so}(t) \equiv \sum_{j\le j_0}\int_{-\infty}^{\infty}|q_j(x,t)|\left(\int_{-\infty}^{x(q_j)}G(\xi)d\xi\right)dx+\sum_{j\ge j_0+1}\int_{-\infty}^{\infty}|q_j(x,t)|\left(\int_{x(q_j)}^{\infty}G(\xi)d\xi\right)dx,\\
H(t)\equiv[1+K_1F((p-1)Ms)]L(t)+K_2[Q_d(t)+E(t)+Q_{so}(t)],\ t\in[(p-1)Ms,pMs),\\
1\le p\le N,
\end{aligned}
$$

where K_1 and K_2 are positive constants to be determined later, and $F(t)=F(u:t)+F(v:t)$ is the modified Glimm functional which was introduced in section 3.

5.2. Basic estimates. In this subsection, we study basic estimates which are necessary for the decay analysis of the nonlinear functional $H(t)$. In the following, the first three lemmas are direct consequences of the smoothness of the shock curves. See [6] and [24].

LEMMA 5.1. *Let $\bar u\in\mathcal{N}$ and $k\in\{1,2,\dots,n\}$. Let us define the states and wave speeds as follows:*

$$
\begin{aligned}
&u=H_k(\xi)(\bar u),\quad u'=H_k(\xi')(u),\quad u''=H_k(\xi+\xi')(\bar u),\\
&\lambda=\lambda_k(\bar u,u),\quad \lambda'=\lambda_k(u,u'),\quad \lambda''=\lambda_k(\bar u,u'').
\end{aligned}
$$

Then we have

$$
\begin{aligned}
|(\xi+\xi')\lambda''-(\xi\lambda+\xi'\lambda')|&=|(\xi+\xi')(\lambda''-\lambda')-\xi(\lambda-\lambda')|\\
&=\mathcal{O}(1)|\xi||\xi'|(|\xi|+|\xi'|).
\end{aligned}
$$

Let us set

$\{\omega_0^+,\omega_1^+,\dots,\omega_n^+\}$: the resolution of a discontinuity $(u(x(\alpha^i)+,t),v(x(\alpha^i)+0,t))$,
$\{\omega_0^-,\omega_1^-,\dots,\omega_n^-\}$: the resolution of a discontinuity $(u(x(\alpha^i)-,t),v(x(\alpha^i)-0,t))$,

$$
q_j^{\pm}(\alpha^i)=(\omega_j^{\pm}-\omega_{j-1}^{\pm})^j,\quad j=1,\dots,n.
$$

LEMMA 5.2. *Suppose that ξ_j,ξ_j', and ξ_j'' satisfy*

$$
H_n(\xi_n)\circ\cdots\circ H_1(\xi_1)(u)=H_n(\xi_n')\circ\cdots\circ H_1(\xi_1')\circ H_n(\xi_n'')\circ\cdots\circ H_1(\xi_1'')(u).
$$

Then, we get

$$
\sum_{i=1}^n|\xi_i-\xi_i'-\xi_i''|=\mathcal{O}(1)\left\{\sum_i|\xi_i'||\xi_i''|(|\xi_i'|+|\xi_i''|)+\sum_{j>i}|\xi_j''||\xi_i'|\right\}.
$$

If the values ξ_i' and ξ are related by

$$
R_i(\xi)(u^*)=H_n(\xi_n')\circ\cdots\circ H_1(\xi_1')(u^*),
$$

then we have

$$
|\xi-\xi_i|+\sum_{j\ne i}|\xi_j'|=\mathcal{O}(1)\left\{|\xi||\xi_i'|(|\xi|+|\xi_i'|)+\sum_{j\ne i}|\xi_j'||\xi|\right\}.
$$

Suppose $\alpha^i=(v_-,v_+)\in J(v)$ is an i-wave in v and u is continuous at $x=x(\alpha^i)$. Recall $\Lambda_p=\{(x,t):-\infty<x<\infty,(p-1)Ms\le t<pMs\},\quad p\in\{1,\dots,N\}$.

Let us set

$$\begin{aligned}
&e(\Lambda_p) \equiv (Q(\Lambda_p) + C(\Lambda_p) + \delta + \epsilon + MsG_0),\\
&\Gamma_s(\alpha^i) \equiv |\alpha^i||q^-(\alpha^i)|(|q^-(\alpha^i)| + |\alpha^i|) \text{ or } |\alpha^i||q^+(\alpha^i)|(|q^+(\alpha^i)| + |\alpha^i|),\\
&\Gamma_d(\alpha^i) \equiv |\alpha^i|\sum_{j>i} |q_j^-(\alpha^i)| \text{ or } |\alpha^i|\sum_{j<i} |q_j^+(\alpha^i)|.
\end{aligned}$$

Then it is easy to see that

$$(\Gamma_s + \Gamma_d)(\alpha^i) = \mathcal{O}(1)\sum_{j=1}^{n} |\alpha^i||q_j^-(\alpha^i)| = \mathcal{O}(1)\sum_{j=1}^{n} |\alpha^i||q_j^+(\alpha^i)|.$$

Remark. If $\alpha^i = (u_-, u_+) \in J(u)$ and v is continuous at $x = x(\alpha^i)$, then we have

$$\begin{aligned}
&\Gamma_s(\alpha^i) \equiv |\alpha^i||q^+(\alpha^i)|(|q^+(\alpha^i)| + |\alpha^i|) \text{ or } |\alpha^i||q^-(\alpha^i)|(|q^-(\alpha^i)| + |\alpha^i|),\\
&\Gamma_d(\alpha^i) \equiv |\alpha^i|\sum_{j<i} |q_j^+(\alpha^i)| \text{ or } |\alpha^i|\sum_{j>i} |q_j^-(\alpha^i)|.
\end{aligned}$$

In the following, we study the variation of $q_j(x,t)$ across the wave.

LEMMA 5.3. *Let $\alpha^i = (v_-, v_+) \in J(v)$ be an i-wave in the time zone Λ_p. Then*

$$q_j^+(\alpha^i) = \begin{cases} q^-(\alpha^i) + [\alpha^i] + \mathcal{O}(1)(\Gamma_s + \Gamma_d)(\alpha^i) + \mathcal{O}(1)|\alpha^i|e(\Lambda_p), & j = i,\\ q_j^-(\alpha^i) + \mathcal{O}(1)(\Gamma_s + \Gamma_d)(\alpha^i) + \mathcal{O}(1)|\alpha^i|e(\Lambda_p), & j \neq i, \end{cases}$$

where $[\alpha^i] = (v_+ - v_-)^i$.

Remark. If $\alpha^i = (u_-, u_+) \in J(u)$, then the same estimates hold by a straightforward calculation.

In the following, we estimates the time-variation of a shock strength as given in Lemma 4.4.

LEMMA 5.4. *Let $\alpha^i(t) = (u_-(t), u_+(t))$, $t \in [(p-1)Ms, pMs)$ be an i-wave issued from $(hr, (p-1)Ms)$ in the simplified wave pattern $u(x,t)$. Then*

$$\frac{d|\alpha^i(t)|}{dt} = \mathcal{O}(1)G(x(\alpha^i))|\alpha^i(t)|,$$

where $\mathcal{O}(1)$ depends only on (1.1).

Proof. The same argument as in Lemma 4.4 holds for this case. □

For a given noninteracting time t, let us denote $J = \{\alpha_i\}_1^m$ by the set of all waves in u and v, and assume that

$$-\infty < x(\alpha_1) < \cdots < 0 \le x(\alpha_k) < x(\alpha_{k+1}) < \cdots < 1 \le x(\alpha_l) < \cdots < x(\alpha_m) < \infty.$$

Without loss of generality, we may assume that $x(\alpha_k(t)) = 0$ and $x(\alpha_l(t)) = 1$.

LEMMA 5.5. *For a given time t and $j \in \{1, 2, \ldots, n\}$,*

$q_j(x,t)$ is differentiable a.e $x \in R$ and $\dfrac{\partial q_j(x,t)}{\partial x} = \mathcal{O}(1)G(x)\displaystyle\sum_{k=1}^{n} q_k(x,t)$, *a.e. $x \in R$.*

Proof. By the construction of a simplified wave pattern, $q_j(x,t)$ is piecewise differentiable in x and $\frac{\partial q_j(x,t)}{\partial x} = 0$ a.e $x \notin [0,1]$. Let $x \in (x(\alpha_i), x(\alpha_{i+1}))$, $i \in \{k, \ldots, l-1\}$. Since $u(x,t)$ and $v(x,t)$ are local steady solutions of (1.1),

$$u_x = (f'(u))^{-1}g(x,u), \quad v_x = (f'(v))^{-1}g(x,v).$$

Therefore, we have

$$v(y) = v(x) + \int_x^y (f'(v))^{-1} g(\xi, v) d\xi, \quad u(y) = u(x) + \int_x^y (f'(u))^{-1} g(\xi, u) d\xi,$$

$$(5.1) v(y) - u(y) = v(x) - u(x) + \int_x^y (f'(v))^{-1} g(\xi, v) - (f'(u))^{-1} g(\xi, u) d\xi.$$

Let us set $h(\xi, p) = f'(p))^{-1} g(\xi, p)$. Then

$$\begin{aligned} &\int_x^y \{(f'(v))^{-1} g(\xi, v) - (f'(u))^{-1} g(\xi, u)\} d\xi = \int_x^y (h(\xi, v) - h(\xi, u)) d\xi \\ &= \int_x^y \left\{ \int_0^1 \frac{\partial}{\partial s} h(\xi, u + s(v-u)) ds \right\} d\xi \\ &= \int_x^y \left\{ \int_0^1 \sum_{i=1}^n \frac{\partial h}{\partial p^i} |_{(\xi, u+s(v-u))} \cdot (v^i - u^i) ds \right\} d\xi \\ &= \mathcal{O}(1) \left(\int_x^y G(\xi) d\xi \right) |v(x) - u(x)|, \end{aligned}$$

where we have used that fact that $\frac{\partial h}{\partial p^i}|_{(\xi, u+s(v-u))} = \mathcal{O}(1) G(\xi)$ and $\sum_1^n |v^i(\xi, t) - u^i(\xi, t)| = \mathcal{O}(1)|v(x,t) - u(x,t)|$. Therefore, in (5.1), we have

$$v(y) - u(y) = v(x) - u(x) + \mathcal{O}(1) \left(\int_x^y G(\xi) d\xi \right) |v(x) - u(x)|.$$

This implies that

$$(v(x) - u(x))_x = \mathcal{O}(1) G(x) \sum_{k=1}^n |q_k(x,t)|.$$

Since $q_i(x,t) = \mathcal{O}(1) l_i(u(x)) \cdot (v(x) - u(x))$, by a direct calculation, we have the following estimate:

$$\frac{\partial q_i(x,t)}{\partial x} = \mathcal{O}(1) G(x) \sum_{k=1}^n |q_k(x,t)|.$$

This completes the proof. □

LEMMA 5.6. *For each $j \in \{1, \dots, n\}$,*

$$\sum_{\alpha^i \in J} \{\lambda(q_j^-(\alpha^i)) |q_j^-(\alpha^i)| - \lambda(q_j^+(\alpha^i)) |q_j^+(\alpha^i)|\} = \mathcal{O}(1) \sum_{k=1}^n \int_0^1 G(x) |q_k(x,t)| dx.$$

Proof. By the construction, the simplified wave pattern is piecewise constant outside an interval $[0,1]$. Therefore,

$$\lambda(q_j^-(\alpha_{i+1})) |q_j^-(\alpha_{i+1})| - \lambda(q_j^+(\alpha_i)) |q_j^+(\alpha_i)| = 0, \quad i \in \{1, \dots, k-1, l, \dots, m-1\},$$

$$\sum_{i=1}^{m-1} \{\lambda(q_j^-(\alpha_{i+1})) |q_j^-(\alpha_{i+1})| - \lambda(q_j^+(\alpha_i)) |q_j^+(\alpha_i)|\}$$

$$
\begin{aligned}
&= \sum_{1}^{k-1} \{\lambda(q_j^-(\alpha_{i+1}))|q_j^-(\alpha_{i+1})| - \lambda(q_j^+(\alpha_i))|q_j^+(\alpha_i)|\} \\
&+ \sum_{k}^{l-1} \{\lambda(q_j^-(\alpha_{i+1}))|q_j^-(\alpha_{i+1})| - \lambda(q_j^+(\alpha_i))|q_j^+(\alpha_i)|\} \\
&+ \sum_{l}^{m-1} \{\lambda(q_j^-(\alpha_{i+1}))|q_j^-(\alpha_{i+1})| - \lambda(q_j^+(\alpha_i))|q_j^+(\alpha_i)|\} \\
&= \sum_{k}^{l-1} \{\lambda(q_j^-(\alpha_{i+1}))|q_j^-(\alpha_{i+1})| - \lambda(q_j^+(\alpha_i))|q_j^+(\alpha_i)|\} \\
&= \sum_{k}^{l-1} II(\alpha_i, \alpha_{i+1}).
\end{aligned}
$$

Let us consider

$$II(\alpha_i, \alpha_{i+1}) = \lambda(q_j^-(\alpha_{i+1}))|q_j^-(\alpha_{i+1})| - \lambda(q_j^+(\alpha_i))|q_j^+(\alpha_i)|,\ i \in \{k, \dots, l-1\}.$$

Case 1. $q_j^+(\alpha_i) \geq 0, q_j^-(\alpha_{i+1}) \geq 0.$

$$
\begin{aligned}
(5.2) \qquad II(\alpha_i, \alpha_{i+1}) &= \lambda(q_j^-(\alpha_{i+1}))|q_j^-(\alpha_{i+1})| - \lambda(q_j^+(\alpha_i))|q_j^+(\alpha_i)| \\
&= f^j(\omega_j(x(\alpha_{i+1})-, t)) - f^j(\omega_{j-1}(x(\alpha_{i+1})-, t)) \\
&- \{f^j(\omega_j(x(\alpha_i)+, t)) - f^j(\omega_{j-1}(x(\alpha_i)+, t))\} \\
&= \int_{x(\alpha_i)}^{x(\alpha_{i+1})} \{f^j(\omega_j(x,t))_x - f^j(\omega_{j-1}(x,t))_x\} dx \\
&= \int_{x(\alpha_i)}^{x(\alpha_{i+1})} \{\nabla_p f^j(\theta_j(x,t)) \cdot (\omega_j(x,t) - \omega_{j-1}(x,t))\}_x dx,
\end{aligned}
$$

where $\theta_j(x,t) = (\theta_j^1(x,t), \dots, \theta_j^n(x,t))$ is on the line segment connecting $\omega_{j-1}(x,t)$ and $\omega_j(x,t)$. By the chain rule, we have

$$
\begin{aligned}
\{\nabla_p f^j(\theta_j(x,t)) \cdot (\omega_j(x,t) - \omega_{j-1}(x,t))\}_x &= (\nabla_p f^j(\theta_j(x,t)))_x \cdot (\omega_j(x,t) - \omega_{j-1}(x,t)) \\
&+ \nabla_p f^j(\theta_j(x,t)) \cdot (\omega_j(x,t) - \omega_{j-1}(x,t))_x.
\end{aligned}
$$

We claim the following:

$$
\begin{aligned}
&\text{(i)} \quad (\nabla_p f^j(\theta_j(x,t)))_x = \mathcal{O}(1)G(x), \\
&\text{(ii)} \quad (\omega_j(x,t) - \omega_{j-1}(x,t))_x = \mathcal{O}(1)G(x) \sum_{k=1}^{n} |q_k(x,t)|.
\end{aligned}
$$

(i) By the chain rule,

$$(\nabla_p f^j(\theta_j(x,t)))_x = \left(\sum_{l=1}^{n} \frac{\partial^2 f^j(\theta_j(x,t))}{\partial p^l \partial p^1} \cdot \theta_j^l(x,t)_x, \dots, \sum_{l=1}^{n} \frac{\partial^2 f^j(\theta_j(x,t))}{\partial p^l \partial p^n} \cdot \theta_j^l(x,t)_x \right).$$

Since $\nabla_p f^j(\theta_j(x,t)) = \mathcal{O}(1)$, $\frac{\partial^2 f^j(\theta_j(x,t))}{\partial p^l \partial p^s} = \mathcal{O}(1)$ and $\theta_j^l(x,t)_x = \mathcal{O}(1)G(x)$, we have

$$(\nabla_p f^j(\theta_j(x,t)))_x = \mathcal{O}(1)G(x).$$

(ii) By definition of $q_j(x,t)$, $\omega_j(x,t) - \omega_{j-1}(x,t) = \mathcal{O}(1)q_j(x,t)$. Therefore, we have

$$(\omega_j(x,t) - \omega_{j-1}(x,t))_x = \mathcal{O}(1)_x q_j(x,t) + \mathcal{O}(1)q_j(x,t)_x.$$

Lemma 5.5 yields

$$(\omega_j(x,t) - \omega_{j-1}(x,t))_x = \mathcal{O}(1)G(x)\sum_{k=1}^{n}|q_k(x,t)|.$$

From the above claim, in (5.2) we have

$$II(\alpha_i,\alpha_{i+1}) = \mathcal{O}(1)\sum_{k=1}^{n}\int_{x(\alpha_i)}^{x(\alpha_{i+1})} G(x)|q_k(x,t)|dx.$$

Case 2. $q_j^+(\alpha_i) \le 0, q_j^-(\alpha_{i+1}) \le 0$. By the same analysis as Case 1, we have

$$II(\alpha_i,\alpha_{i+1}) = \mathcal{O}(1)\sum_{k=1}^{n}\int_{x(\alpha_i)}^{x(\alpha_{i+1})} G(x)|q_k(x,t)|dx.$$

Case 3. $q_j^+(\alpha_i) > 0$, $q_j^-(\alpha_{i+1}) < 0$. Since $q_j(x,t)$ is continuous on $[x(\alpha_i), x(\alpha_{i+1})]$, by the intermediate value theorem, $q_j(x_0(t),t) = 0$ for some $x_0(t) \in (x(\alpha_i), x(\alpha_{i+1}))$. Using this fact, one has

$$|q_j^+(\alpha_i)| = |q_j(x_0) - q_j^+(\alpha_i)| = \left|\int_{x(\alpha_i)}^{x_0} q_j(x,t)_x dx\right| = \mathcal{O}(1)\sum_{k=1}^{n}\int_{x(\alpha_i)}^{x_0} G(x)|q_k(x,t)|dx.$$

On the other hand,

$$\begin{aligned}|q_j^-(\alpha_{i+1})| = |q_j^-(\alpha_{i+1}) - q_j(x_0)| &= \left|\int_{x_0}^{x(\alpha_{i+1})} q_j(x,t)_x dx\right| \\ &= \mathcal{O}(1)\sum_{k=1}^{n}\int_{x_0}^{x(\alpha_{i+1})} G(x)|q_k(x,t)|dx.\end{aligned}$$

Since $|\lambda(q_j^-(\alpha_{i+1}))| = \mathcal{O}(1)$, $|\lambda(q_j^+(\alpha_i))| = \mathcal{O}(1)$,

$$II(\alpha_i,\alpha_{i+1}) = \mathcal{O}(1)\sum_{k=1}^{n}\int_{x(\alpha_i)}^{x(\alpha_{i+1})} G(x)|q_k(x,t)|dx.$$

Case 4. $q_j^+(\alpha_i) < 0, q_j^-(\alpha_{i+1}) > 0$. By the same analysis as Case 3, we have

$$II(\alpha_i,\alpha_{i+1}) \le \mathcal{O}(1)\sum_{k=1}^{n}\int_{x(\alpha_i)}^{x(\alpha_{i+1})} G(x)|q_k(x,t)|dx.$$

This completes the proof. □

5.3. L^1 stability and uniqueness. Recall that the open interval $I_p = ((p-1)Ms, pMs), p \in \{1, \dots, N\}$ is the union of two disjoint sets $I_p = I_p^1 \cup I_p^2$, where I_p^1 is the set of all countable interaction times such that $H(t)$ is simply continuous, and I_p^2 is the set of all differentiable points of $H(t)$.

Lemma 5.7. *The nonlinear functional $H(t)$ is "almost decreasing" for the simplified wave patterns in the sense that*

$$H(pMs+) \leq H((p-1)Ms+) + \mathcal{O}(1)e(\Lambda_p)Ms,$$

where $\mathcal{O}(1)$ depends only on (1.1) *and $e(\Lambda_p) = Q(\Lambda_p) + C(\Lambda_p) + (\epsilon + \delta + MsG_0)$.*

Proof. We will study the time rate of a change for each component functional separately. Set

$$\Gamma \equiv \Gamma_s + \Gamma_d + \Gamma_{so}, \qquad \Gamma_s \equiv \sum_\beta \Gamma_s(\beta),$$
$$\Gamma_d \equiv \sum_\beta \Gamma_d(\beta), \qquad \Gamma_{so} \equiv \sum_1^n \int_0^1 G(x)|q_j(x,t)|dx.$$

In the following proof, we will use a similar analysis to that in Lemma 5.1 of [24].

Step 1. $\frac{dL(t)}{dt} \leq \mathcal{O}(1)\Gamma + \mathcal{O}(1)(T.V + G_1)e(\Lambda_p), t \in I_p^2$. By definition of $L(t)$,

$$\frac{dL(t)}{dt} = \sum_{j=1}^n \frac{dL_j(t)}{dt} = \sum_{j=1}^n \frac{d}{dt}\int_{-\infty}^\infty |q_j(x,t)|dx.$$

Let $j \in \{1, 2, \dots, n\}$. Then it follows from Lemma 5.6 that

$$\begin{aligned}(5.3) \qquad & \sum_{\alpha^i \in J} \{\lambda(q_j^-(\alpha^i))|q_j^-(\alpha^i)| - \lambda(q_j^+(\alpha^i))|q_j^+(\alpha^i)|\} \\ &= \sum_{i=1}^m \{\lambda(q_j^-(\alpha_i))|q_j^-(\alpha_i)| - \lambda(q_j^+(\alpha_i))|q_j^+(\alpha_i)|\} \\ &= \mathcal{O}(1)\sum_{l=1}^n \int_0^1 G(x)|q_l(x,t)|dx.\end{aligned}$$

Let us consider $\frac{d}{dt}\int_{-\infty}^\infty |q_j(x,t)|dx$; then by a direct calculation, we have

$$\begin{aligned}(5.4) \qquad & \frac{d}{dt}\int_{-\infty}^\infty |q_j(x,t)|dx = \sum_{\alpha^i \in J} \dot{x}(\alpha^i)(|q_j^-(\alpha^i)| - |q_j^+(\alpha^i)|) \\ &= \sum_{\alpha^i \in J} \{(\dot{x}(\alpha^i) - \lambda(q_j^-(\alpha^i)))|q_j^-(\alpha^i)| - (\dot{x}(\alpha^i) - \lambda(q_j^+(\alpha^i)))|q_j^+(\alpha^i)|\} \\ &+ \sum_{\alpha^i \in J} \{\lambda(q_j^-(\alpha^i))|q_j^-(\alpha^i)| - \lambda(q_j^+(\alpha^i))|q_j^+(\alpha^i)|\} \\ &= \sum_{\alpha^i \in J} I_j(\alpha^i) + \mathcal{O}(1)\sum_{l=1}^n \int_0^1 G(x)|q_l(x,t)|dx \quad \text{by (5.3)}.\end{aligned}$$

Let us set $I_j(\alpha^i) \equiv (\dot{x}(\alpha^i) - \lambda(q_j^-(\alpha^i)))|q_j^-(\alpha^i)| - (\dot{x}(\alpha^i) - \lambda(q_j^+(\alpha^i)))|q_j^+(\alpha^i)|$. Then, it follows from [24] that

$$I_j(\alpha^i) = \mathcal{O}(1)(\Gamma_s(\alpha^i) + \Gamma_d(\alpha^i)) + \mathcal{O}(1)|\alpha^i|e(\Lambda_p).$$

From (5.4), we get

$$\frac{d}{dt}\int_{-\infty}^{\infty}|q_j(x,t)|dx \le \mathcal{O}(1)\sum_{\beta\in J}(\Gamma_s+\Gamma_d)(\beta)+\mathcal{O}(1)(T.V+G_1)e(\Lambda_p)$$
$$+\,\mathcal{O}(1)\sum_{l=1}^{n}\int_0^1 G(x)|q_j(x,t)|dx.$$

Therefore, we have

$$\frac{dL(t)}{dt}\le \mathcal{O}(1)\Gamma+\mathcal{O}(1)(T.V+G_1)e(\Lambda_p).$$

Step 2. $\frac{dQ_d(t)}{dt}\le -\lambda_0\Gamma_d+\mathcal{O}(1)(T.V+G_1)\Gamma+\mathcal{O}(1)\left(\sum_\alpha G(x(\alpha))|\alpha|\right)L(t)$ $+\mathcal{O}(1)(T.V+G_1)e(\Lambda_p)$. By definition of $Q_d(t)$,

$$\text{(5.5)}\quad \frac{dQ_d(\alpha^i)}{dt}=|\alpha^i|\left\{\sum_{j>i}\frac{d}{dt}\int_{-\infty}^{x(\alpha^i)}|q_j(x,t)|dx+\sum_{j<i}\frac{d}{dt}\int_{x(\alpha^i)}^{\infty}|q_j(x,t)|dx\right\}$$
$$+\,\frac{d|\alpha^i|}{dt}\left\{\sum_{j>i}\int_{-\infty}^{x(\alpha^i)}|q_j(x,t)|dx+\sum_{j<i}\int_{x(\alpha^i)}^{\infty}|q_j(x,t)|dx\right\}$$
$$\le |\alpha^i|\left\{\sum_{j>i}\frac{d}{dt}\int_{-\infty}^{x(\alpha^i)}|q_j(x,t)|dx+\sum_{j<i}\frac{d}{dt}\int_{x(\alpha^i)}^{\infty}|q_j(x,t)|dx\right\}$$
$$+\,\mathcal{O}(1)G(x(\alpha^i))|\alpha^i|\sum_{j\ne i}L_j(t).$$

By the strict hyperbolicity of (1.1), for some positive constant λ_0,

$$\begin{aligned}&\lambda_j(u)-\lambda(\alpha^i)>\lambda_0,\quad j>i,\\&\lambda_j(u)-\lambda(\alpha^i)<-\lambda_0,\quad j<i.\end{aligned}$$

Then by the same calculation as in [24], from (5.5) we have the following estimate:

$$\frac{dQ_d(\alpha^i)}{dt}\le -\lambda_0\Gamma_d(\alpha^i)+\mathcal{O}(1)|\alpha^i|\Gamma+\mathcal{O}(1)|\alpha^i|(T.V+G_1)e(\Lambda_p)$$
$$+\,\mathcal{O}(1)G(x(\alpha))|\alpha^i|\sum_{j\ne i}L_j(t).$$

Hence, we have

$$\frac{dQ_d(t)}{dt}\le -\lambda_0\sum_{\beta\in J}\Gamma_d(\beta)+\mathcal{O}(1)(T.V+G_1)\Gamma+\mathcal{O}(1)(T.V+G_1)e(\Lambda_p)$$
$$+\,\mathcal{O}(1)\left(\sum_\alpha G(x(\alpha))|\alpha|\right)L(t).$$

Step 3. $\frac{dE(t)}{dt}\le -C_4\Gamma_s+\mathcal{O}(1)(T.V+G_1)\Gamma+\mathcal{O}(1)(T.V+G_1)e(\Lambda_p)$ $+\mathcal{O}(1)(\sum_\alpha G(x(\alpha))|\alpha|)L(t)$, where C_4 is a positive constant depending only on (1.1). By definition of $E(t)$,

$$\frac{dE(t)}{dt}=\sum_{\alpha^i\in J}\frac{d}{dt}\left[|\alpha^i|\left\{\int_{-\infty}^{x(\alpha^i)}(q_i(x,t))_-dx+\int_{x(\alpha^i)}^{\infty}(q_i(x,t))_+dx\right\}\right]$$

$$= \sum_{\alpha^i \in J} \frac{dE(\alpha^i)}{dt}.$$

Let us assume that $\alpha^i \in J(v)$, and u is continuous at $x = x(\alpha^i)$. The other case ($\alpha^i \in J(u)$) is treated similarly.

$$\begin{aligned}
\frac{dE(\alpha^i)}{dt} &= \frac{d|\alpha^i|}{dt}\left\{\int_{-\infty}^{x(\alpha^i)} (q_i(x,t))_- dx + \int_{x(\alpha^i)}^{\infty} (q_i(x,t))_+ dx\right\} \\
&\quad + |\alpha^i|\left\{\frac{d}{dt}\int_{-\infty}^{x(\alpha^i)} (q_i(x,t))_- dx + \frac{d}{dt}\int_{x(\alpha^i)}^{\infty} (q_i(x,t))_+ dx\right\} \\
&= \mathcal{O}(1)G(x(\alpha^i))|\alpha^i|\left\{\int_{-\infty}^{x(\alpha^i)} (q_i(x,t))_- dx + \int_{x(\alpha^i)}^{\infty} (q_i(x,t))_+ dx\right\} \\
&\quad + |\alpha^i|\left\{\frac{d}{dt}\int_{-\infty}^{x(\alpha^i)} (q_i(x,t))_- dx + \frac{d}{dt}\int_{x(\alpha^i)}^{\infty} (q_i(x,t))_+ dx\right\} \\
&\le \mathcal{O}(1)G(x(\alpha^i))|\alpha^i|L_i(t) \\
&\quad + |\alpha^i|\left\{\frac{d}{dt}\int_{-\infty}^{x(\alpha^i)} (q_i(x,t))_- dx + \frac{d}{dt}\int_{x(\alpha^i)}^{\infty} (q_i(x,t))_+ dx\right\},
\end{aligned}$$

where we have used the fact that $\frac{d|\alpha^i|}{dt} = \mathcal{O}(1)G(x(\alpha^i))|\alpha^i|$. By the same analysis as in Lemma 5.1 of [24], we have

$$\begin{aligned}
\frac{dE(\alpha^i)}{dt} &\le -C_4|\alpha^i|(|q^-(\alpha^i)| + |\alpha^i|)|q^-(\alpha^i)| + \mathcal{O}(1)(|\alpha^i|\Gamma(\alpha^i) + |\alpha^i|e(\Lambda_p)) \\
&\quad + \mathcal{O}(1)|\alpha^i|\left\{\sum_{\beta\in J}\Gamma(\beta) + (T.V + G_1)e(\Lambda_p) + \sum_{j=1}^{n}\int_0^1 G(x)|q_j(x,t)|dx\right\} \\
&\quad + \mathcal{O}(1)G(x(\alpha^i))|\alpha^i|L_i(t).
\end{aligned}$$

Hence, we have

$$\frac{dE(t)}{dt} \le -C_4\Gamma_s + \mathcal{O}(1)(T.V + G_1)\Gamma + \mathcal{O}(1)\left(\sum_{\alpha} G(x(\alpha))|\alpha|\right)L(t)$$

$$+\mathcal{O}(1)(T.V + G_1)e(\Lambda_p),$$

where C_4 is a positive constant depending only on (1.1).

Step 4. $\frac{dQ_{so}(t)}{dt} \le -\lambda_0\Gamma_{so} + \mathcal{O}(1)G_1\{\Gamma + (T.V + G_1)e(\Lambda_p)\}$.

By definition of $Q_{so}(t)$, we have

$$\begin{aligned}
Q_{so}(t) &= \sum_{j\le j_0}\int_{-\infty}^{\infty}|q_j(x,t)|\left(\int_{-\infty}^{x(q_j)} G(\xi)d\xi\right)dx \\
&\quad + \sum_{j\ge j_0+1}\int_{-\infty}^{\infty}|q_j(x,t)|\left(\int_{x(q_j)}^{\infty} G(\xi)d\xi\right)dx \\
&= \sum_{j\le j_0} I_j + \sum_{j\ge j_0+1} II_j.
\end{aligned}$$

For a given t, let us denote the locations of waves as follows:

$$\begin{aligned}-\infty < x(\alpha_1) < \cdots < x(\alpha_k) &= 0 < x(\alpha_{k+1}) < \cdots < x(\alpha_l)\\ &= 1 < x(\alpha_{l+1}) < \cdots < x(\alpha_m) < \infty.\end{aligned}$$

For notational convenience, let us denote $x(\alpha_0) \equiv -\infty$ and $x(\alpha_{m+1}) \equiv \infty$.

Case 1. $j \le j_0$.

Since $\int_{-\infty}^0 G(\xi)d\xi = 0$,

$$\begin{aligned}I_j &= \int_{-\infty}^{\infty} |q_j(x,t)| \left(\int_{-\infty}^{x(q_j)} G(\xi)d\xi\right) dx\\ &= \int_{x(\alpha_k)}^{x(\alpha_{k+1})} |q_j(x,t)| \left(\int_{-\infty}^{x(q_j)} G(\xi)d\xi\right) dx + \int_{x(\alpha_{k+1})}^{x(\alpha_{k+2})} |q_j(x,t)| \left(\int_{-\infty}^{x(q_j)} G(\xi)d\xi\right) dx\\ &+ \cdots + \int_{x(\alpha_{l-1})}^{x(\alpha_l)} |q_j(x,t)| \left(\int_{-\infty}^{x(q_j)} G(\xi)d\xi\right) dx + G_1 \int_{x(\alpha_l)}^{x(\alpha_{l+1})} |q_j(x,t)|dx\\ &+ \cdots + G_1 \int_{x(\alpha_{m-1})}^{x(\alpha_m)} |q_j(x,t)|dx.\end{aligned}$$

Then by a direct calculation, we have

$$\begin{aligned}\frac{dI_j}{dt} &\le \sum_{i=k}^{l-1} \left(\int_{-\infty}^{x(q_j(\alpha_i))} G(\xi)d\xi\right) \dot{x}(\alpha_i)(|q_j^-(\alpha_i)| - |q_j^+(\alpha_i)|)\\ &+ G_1 \sum_{i=l}^{m} \dot{x}(\alpha_i)(|q_j^-(\alpha_i)| - |q_j^+(\alpha_i)|) + \int_0^1 |q_j(x,t)|\dot{x}(q_j)G(x(q_j))dx\\ &\le \mathcal{O}(1)G_1 \left\{\sum_{i=k}^{m} (\Gamma_s + \Gamma_d)(\alpha_i) + \Gamma_{so} + (T.V + G_1)e(\Lambda_p)\right\} - \lambda_0 \int_0^1 G(x)|q_j(x,t)|dx.\end{aligned}$$

In the above calculation, we have used $\dot{x}(q_j) \le -\lambda_0$. Hence, we have

$$\begin{aligned}\frac{dI}{dt} = \sum_{j\le j_0} \frac{dI_j}{dt} &\le j_0 \mathcal{O}(1)G_1 \left\{\sum_{i=k}^{m} (\Gamma_s + \Gamma_d)(\alpha_i) + \Gamma_{so} + (T.V + G_1)e(\Lambda_p)\right\}\\ &- \lambda_0 \sum_{j\le j_0} \int_0^1 G(x)|q_j(x,t)|dx\\ &\le \mathcal{O}(1)G_1 \left\{\sum_{i=k}^{m} (\Gamma_s + \Gamma_d)(\alpha_i) + \Gamma_{so} + (T.V + G_1)e(\Lambda_p)\right\}\\ &- \lambda_0 \sum_{j\le j_0} \int_0^1 G(x)|q_j(x,t)|dx.\end{aligned}$$

Case 2. $j \ge j_0 + 1$. Since $\int_1^\infty G(\xi)d\xi = 0$,

$$\begin{aligned}II_j &= \int_{-\infty}^{\infty} |q_j(x,t)| \left(\int_{x(q_j)}^{\infty} G(\xi)d\xi\right) dx\\ &= G_1 \int_{x(\alpha_1)}^{x(\alpha_2)} |q_j(x,t)|dx + \cdots + G_1 \int_{x(\alpha_{k-1})}^{x(\alpha_k)} |q_j(x,t)|dx\end{aligned}$$

$$+\int_{x(\alpha_k)}^{x(\alpha_{k+1})} |q_j(x,t)| \left(\int_{x(q_j)}^{\infty} G(\xi)d\xi\right) dx + \cdots + \int_{x(\alpha_{l-1})}^{x(\alpha_l)} |q_j(x,t)| \left(\int_{x(q_j)}^{\infty} G(\xi)d\xi\right) dx.$$

Then by a direct calculation, we have

$$\begin{aligned}
&\frac{dII_j}{dt} = G_1 \sum_{i=1}^{k-1} \dot{x}(\alpha_i)(|q_j^-(\alpha_i)| - |q_j^+(\alpha_i)|) \\
&+ \sum_k^l \left(\int_{x(\alpha_i)}^{\infty} G(\xi)d\xi\right) \dot{x}(\alpha_i)(|q_j^-(\alpha_i)| - |q_j^+(\alpha_i)|) + \int_0^1 |q_j(x,t)|(-\dot{x}(q_j))G(x(q_j))dx \\
&\leq \mathcal{O}(1)G_1 \left\{\sum_{i=1}^l (\Gamma_s + \Gamma_d)(\alpha_i) + \Gamma_{so} + (T.V + G_1)e(\Lambda_p)\right\} - \lambda_0 \int_0^1 G(x)|q_j(x,t)|dx.
\end{aligned}$$

In the above calculations, we have used the fact that $\dot{x}(q_j) \geq \lambda_0$. Hence we have

$$\begin{aligned}
\frac{dII}{dt} \leq \sum_{j\geq j_0+1} \frac{dII_j}{dt} &= \mathcal{O}(1)G_1(n-j_0)\sum_1^l \{(\Gamma_s + \Gamma_d)(\alpha_i) + \Gamma_{so} \\
&+ (T.V + G_1)e(\Lambda_p)\} - \lambda_0 \sum_{j\geq j_0+1} \int_0^1 G(x)|q_j(x,t)|dx.
\end{aligned}$$

By combining Cases 1 and 2, we have

$$\frac{dQ_{so}(t)}{dt} \leq \mathcal{O}(1)G_1\{\Gamma + (T.V + G_1)e(\Lambda_p)\} - \lambda_0\Gamma_{so}.$$

Step 5. Let us choose $\tilde{c}$ such that $0 < \tilde{c} < \min\{\lambda_0, C_4\}$, where C_4 is the positive constant in Step 3. From definition of $H(t)$ and Steps 1–4, for $t \in I_p^2$, we have

$$\begin{aligned}
\frac{dL(t)}{dt} &\leq \mathcal{O}(1)\Gamma + \mathcal{O}(1)(T.V + G_1)e(\Lambda_p), \\
\frac{dQ_d(t)}{dt} &\leq -\tilde{c}\Gamma_d + \mathcal{O}(1)(T.V + G_1)\Gamma + \mathcal{O}(1)\left(\sum G(x(\alpha))|\alpha|\right)L(t) \\
&\quad + \mathcal{O}(1)(T.V + G_1)e(\Lambda_p), \\
\frac{dE(t)}{dt} &\leq -\tilde{c}\Gamma_s + \mathcal{O}(1)(T.V + G_1)\Gamma + \mathcal{O}(1)\left(\sum G(x(\alpha))|\alpha|\right)L(t) \\
&\quad + \mathcal{O}(1)(T.V + G_1)e(\Lambda_p), \\
\frac{dQ_{so}(t)}{dt} &\leq -\tilde{c}\Gamma_{so} + \mathcal{O}(1)G_1\Gamma + \mathcal{O}(1)G_1(T.V + G_1)e(\Lambda_p),
\end{aligned}$$

$$\begin{aligned}
\frac{dH(t)}{dt} &= (1 + K_1F((p-1)Ms))\frac{dL(t)}{dt} + K_2\left(\frac{dQ_d(t)}{dt} + \frac{dE(t)}{dt} + \frac{dQ_{so}(t)}{dt}\right) \\
&\leq [\mathcal{O}(1)\{1 + K_1F((p-1)Ms)\} + \mathcal{O}(1)K_2(T.V + G_1) + \mathcal{O}(1)K_2G_1 - \tilde{c}K_2]\Gamma \\
&+ [\mathcal{O}(1)(T.V + G_1)(1 + K_1F((p-1)Ms)) + \mathcal{O}(1)K_2(T.V + G_1) \\
&+ \mathcal{O}(1)K_2G_1(T.V + G_1)]e(\Lambda_p) + \mathcal{O}(1)K_2\left(\sum G(x(\alpha))|\alpha|\right)L(t).
\end{aligned} \tag{5.6}$$

Since $L(t)$ is Lipschitz continuous on $((p-1)Ms, pMs)$, we have

$$L(t) \leq \mathcal{O}(1)Ms + L(pMs-). \tag{5.7}$$

By definition of $H(t)$, there is a jump across the $t = pMs, p \in \{1, \dots, N\}$. Next, we estimate the size of this jump.

$$\begin{aligned} H(pMs+) - H(pMs-) = {} & [(1 + K_1 F(pMs))L(pMs+) \\ & + K_2(Q_d(pMs+) + E(pMs+) + Q_{so}(pMs+))] \\ & - [(1 + K_1 F((p-1)Ms)))L(pMs-) + K_2(Q_d(pMs-) \\ & + E(pMs-) + Q_{so}(pMs-))] = \sum_{i=1}^{5} I_i, \end{aligned}$$

where

$$\begin{aligned} I_1 &\equiv K_1(F(pMs) - F((p-1)Ms))L(pMs-), \\ I_2 &\equiv (1 + K_1 F(pMs))(L(pMs+) - L(pMs-)), \\ I_3 &\equiv K_2(Q_d(pMs+) - Q_d(pMs-)), \\ I_4 &\equiv K_2(E(pMs+) - E(pMs-)), \\ I_5 &\equiv K_2(Q_{so}(pMs+) - Q_{so}(pMs-)). \end{aligned}$$

By Lemma 3.2

$$F(pMs) - F((p-1)Ms) \leq -\frac{1}{2}\left(Q(\Lambda_p) + C(\Lambda_p)\right).$$

Therefore, we have

$$I_1 \leq -\frac{K_1}{2}\left(Q(\Lambda_p) + C(\Lambda_p)\right) L(pMs-). \tag{5.8}$$

On the other hand, the difference of a wave pattern at time $t = pNs+$ and $t = pMs-$ is due to interactions, cancellations, and errors by the scheme, so we have

$$L(pMs+) - L(pMs-) \leq \mathcal{O}(1)e(\Lambda_p)Ms.$$

Hence,

$$I_2 \leq \mathcal{O}(1)(1 + K_1 F(pMs))e(\Lambda_p)Ms. \tag{5.9}$$

By definition of $Q_d(t)$, I_3 can be estimated by considering the following two terms: one term is the product of change of the wave strengths and the L^1 norm at $t = pMs-$, and the other term is the product of change of the L_1 norm times the wave strengths. Therefore, we have

$$I_3 \leq C_5 K_2(T.V + G_1)e(\Lambda_p)Ms + C_5 K_2(Q(\Lambda_p) + C(\Lambda_p))L(pMs-). \tag{5.10}$$

By the same argument as above, we have

$$I_4 \leq C_5 K_2(T.V + G_1)e(\Lambda_p)Ms + C_5 K_2(Q(\Lambda_p) + C(\Lambda_k))L(pMs-). \tag{5.11}$$

Similarly, we have

$$Q_{so}(pMs+) - Q_{so}(pMs-) \leq \mathcal{O}(1)G_1 e(\Lambda_p)Ms,$$

i.e.,

$$I_5 \leq C_5 K_2 G_1 e(\Lambda_p)Ms. \tag{5.12}$$

Summing up all I_k's (5.8)–(5.12), we have

$$\begin{aligned}(5.13)\quad H(pMs+) - H(pMs-) &\leq \left(2C_5K_2 - \frac{K_1}{2}\right)(Q(\Lambda_p) + C(\Lambda_p))L(pMs-)\\ &+ [\mathcal{O}(1)(1+K_1F(pMs)) + \mathcal{O}(1)K_2(T.V + G_1) + \mathcal{O}(1)K_2G_1]\, e(\Lambda_p)Ms.\end{aligned}$$

If we integrate (5.6) from $(p-1)Ms$ to pMs, then by using (5.7) we have

$$\begin{aligned}(5.14)\qquad H(pMs-) - H((p-1)Ms+) &\leq [\mathcal{O}(1)(1+K_1F((p-1)Ms))\\ &+ \mathcal{O}(1)K_2(T.V + G_1) + \mathcal{O}(1)K_2G_1 - \tilde{c}K_2]\int \Gamma(t)dt\\ &+ [\mathcal{O}(1)(T.V + G_1)(1+K_1F((p-1)Ms)) + \mathcal{O}(1)K_2(T.V+G_1)\\ &+ \mathcal{O}(1)K_2G_1(T.V+G_1)]e(\Lambda_p)Ms + \mathcal{O}(1)K_2Q_1(\Lambda_p)Ms\\ &+ \mathcal{O}(1)K_2Q_1(\Lambda_p)L(pMs-),\end{aligned}$$

where the integral is over $((p-1)Ms, pMs)$, and we have used the fact that

$$\int \sum_{\alpha} G(x(\alpha))|\alpha| dt = \mathcal{O}(1)Q_1(\Lambda_p).$$

From (5.13) and (5.14), we have

$$\begin{aligned}H(pMs+) - H((p-1)Ms+) &\leq [\mathcal{O}(1)(1+K_1F((p-1)Ms))\\ &+ \mathcal{O}(1)K_2(T.V+G_1) + \mathcal{O}(1)K_2G_1 - \tilde{c}K_2]\int \Gamma(t)dt\\ &+ [\mathcal{O}(1)(T.V+G_1)(1+K_1F((p-1)Ms)) + \mathcal{O}(1)K_2(T.V+G_1)\\ &+ \mathcal{O}(1)K_2G_1(T.V+G_1) + \mathcal{O}(1)(1+K_1F(pMs+))\\ &+ \mathcal{O}(1)K_2G_1 + \mathcal{O}(1)K_2]e(\Lambda_p)Ms\\ &+ \left[2C_5K_2 + \mathcal{O}(1)K_2 - \frac{K_1}{2}\right](Q(\Lambda_p) + C(\Lambda_p))L(pMs-).\end{aligned}$$

Since $F(t), G_0, G_1$, and $T.V$ are sufficiently small, we can choose positive constants K_1 and K_2 so that

$$\begin{aligned}&\mathcal{O}(1)(1+K_1F((p-1)Ms)) + \mathcal{O}(1)K_2(T.V+G_1) + \mathcal{O}(1)K_2G_1 - \tilde{c}K_2 < 0,\\ &2C_5K_2 + \mathcal{O}(1)K_2 - \frac{K_1}{2} < 0.\end{aligned}$$

Then for such K_1 and K_2, we have

$$H(pMs+) \leq H((p-1)Ms+) + \mathcal{O}(1)e(\Lambda_p)Ms.$$

This completes the proof. □

Using Lemma 5.7 successively, we obtain the following estimate.

LEMMA 5.8. *Let $\bar{u}_r(x,t)$ and $\bar{v}_r(x,t)$ be two simplified wave patterns of* (1.1) *corresponding to initial data $u_0(x)$ and $v_0(x)$, respectively. If $u_0(x) - v_0(x) \in L^1(R)$, then we have*

$$H(T) \leq H(0) + \mathcal{O}(1)(Q(\Lambda_T) + C(\Lambda_T))Ms + \mathcal{O}(1)(\epsilon + \delta + MsG_0)T.$$

Proof. Let $\bar{u}_r(x,t)$ and $\bar{v}_r(x,t)$ be the simplified wave patterns and $T = NMs$. By Lemma 5.7, we have

$$H(NMs+) \leq H((N-1)Ms+) + \mathcal{O}(1)e(\Lambda_N)Ms.$$

If we use Lemma 5.7 successively in p, we obtain

$$H(T) \leq H(0) + \mathcal{O}(1)(Q(\Lambda_T) + C(\Lambda_T))Ms + \mathcal{O}(1)(\epsilon + \delta + MsG_0)T.$$

This completes the proof. □

Let us define the nonlinear functional $H(t) = H[u(\cdot,t), v(\cdot,t)]$ for two Glimm solutions $u(x,t)$ and $v(x,t)$ by

$$H[u(\cdot,t), v(\cdot,t)] = \lim_{r,\epsilon,\delta\to 0} H[\bar{u}_r(\cdot,t), \bar{v}_r(\cdot,t)],$$

where $\bar{u}_r(x,t)$ and $\bar{v}_r(x,t)$ are the simplified wave patterns of $u(x,t)$ and $v(x,t)$, respectively. Next, we establish L^1 stability as a direct consequence of the above lemma.

Theorem 5.9. *Let $u(x,t)$ and $v(x,t)$ be two weak solutions corresponding to initial data $u_0(x)$ and $v_0(x)$, respectively. If $u_0(x) - v_0(x) \in L^1(R)$, then we have*

$$H(t) \leq H(0),$$
$$\|u(\cdot,t) - v(\cdot,t)\|_{L^1(R)} \leq C\|u_0(\cdot) - v_0(\cdot)\|_{L^1(R)} \quad \textit{for } t \geq 0,$$

where C is independent of t.

Proof. (1) Let $\bar{u}_r(x,t)$ and $\bar{v}_r(x,t)$ be two simplified wave patterns of (1.1) such that

$$\lim_{r,\epsilon,\delta\to 0} \bar{u}_r(x,t) = u(x,t), \qquad \lim_{r,\epsilon,\delta\to 0} \bar{v}_r(x,t) = v(x,t) \quad \text{in } L^1_{loc}(R \times R_+).$$

Since $H[u(\cdot,t), v(\cdot,t)] = \lim_{r,\epsilon,\delta\to 0} H[\bar{u}_r, \bar{v}_r]$, it follows from Lemma 5.8 that

$$H(t) \leq H(0).$$

(2) Since $H[u(\cdot,t), v(\cdot,t)]$ is equivalent to $\|u(\cdot,t) - v(\cdot,t)\|_{L^1(R)}$ (see section 5.1), i.e.,

$$\frac{1}{C_3}\|u(\cdot,t) - v(\cdot,t)\|_{L^1(R)} \leq H(t) \leq 2C_3\|u(\cdot,t) - v(\cdot,t)\|_{L^1(R)}$$

for some positive constant C_3. Therefore, we have

$$\|u(\cdot,t) - v(\cdot,t)\|_{L^1(R)} \leq C_3 H(t) \leq C_3 H(0) \leq 2C_3^2\|u_0(x) - v_0(x)\|_{L^1(R)}.$$

Let us set $C = 2C_3^2$; then we have the desired result. □

Remark. As an immediate consequence of Theorem 5.9, we have the uniqueness of the Glimm solutions.

Acknowledgments. The author wishes to thank to Professor Tai-Ping Liu and Professor Tong Yang for helpful discussions.

REFERENCES

[1] D. Amadori and G. Guerra, *Uniqueness and continuous dependence for systems of balanced laws with dissipation*, Nonlinear Anal., submitted.

[2] A. Bressan, *A locally contractive metrics for systems of conservation laws*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 22 (1995), pp. 109–135.

[3] A. Bressan and R. M. Colombo, *The semigroup generated by* 2×2 *conservation laws*, Arch. Ration. Mech. Anal., 133 (1995), pp. 1–75.

[4] A. Bressan, G. Crasta, and B. Piccoli, *Well Posedness of the Cauchy Problem for $n \times n$ Systems of Conservation Laws*, Mem. Amer. Math. Soc., 146 (2000).

[5] A. Bressan and P. LeFloch, *Uniqueness of weak solutions to systems of conservation laws*, Arch. Ration. Mech. Anal., 140 (1997), pp. 301–317.

[6] A. Bressan, T.-P. Liu, and T. Yang, *L^1 stability estimates for $n \times n$ conservation laws*, Arch. Ration. Mech. Anal., 149 (1999), pp. 1–22.

[7] G. Q. Chen and J. Glimm, *Global solutions to the compressible Euler equations with geometrical structure*, Comm. Math. Phys., 180 (1996), pp. 153–193.

[8] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves*, Interscience, New York, 1948.

[9] G. Crasta and B. Piccoli, *Viscosity solutions and uniqueness for systems of inhomogeneous balance laws*, Discrete Contin. Dynam. Systems, 3 (1997), pp. 477–502.

[10] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.

[11] J. Glimm and P. D. Lax, *Decay of Solutions of Systems of Hyperbolic Conservation Laws*, Mem. Amer. Math. Soc. 101, AMS, Providence, RI, 1970.

[12] A. L. Hoffman, *A single fluid model for shock formation in MHD shock tubes*, J. Plasma. Phys., 1 (1967), pp. 192–207.

[13] B. Keyfitz, *Solutions with shocks: An example of L_1-contractive semigroup*, Comm. Pure Appl. Math., 24 (1971), pp. 125–132.

[14] S. N. Krushkov, *Generalized solutions of the Cauchy problem in the large for nonlinear equations of first order*, Dokl. Akad. Nauk, 187 (1969), pp. 29–32 (in Russian).

[15] P. D. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[16] P. D. Lax, *Shock waves and entropy*, in Contributions to Nonlinear Functional Analysis, E. Zarantonello, ed., Academic Press, New York, 1971, pp. 603–634.

[17] W.-C. Lien, *Hyperbolic conservation laws with a moving source*, Comm. Pure Appl. Math., 52 (1999), pp. 1075–1098.

[18] T.-P. Liu, *The deterministic version of the Glimm scheme*, Comm. Math. Phys., 57 (1975), pp. 135–148.

[19] T.-P. Liu, *Quasilinear hyperbolic systems*, Comm. Math. Phys., 68 (1979), pp. 141–172.

[20] T.-P. Liu, *Nonlinear stability and instability of transonic gas flows through a nozzle*, Comm. Math. Phys., 83 (1982), pp. 243–260.

[21] T.-P. Liu, *Nonlinear resonance for quasilinear hyperbolic equation*, J. Math. Phys., 28 (1987), pp. 2593–2602.

[22] T.-P. Liu and T. Yang, *A new entropy functional for a scalar conservation law*, Comm. Pure Appl. Math., 52 (1999), pp. 1427–1442.

[23] T.-P. Liu and T. Yang, *L^1 stability for 2×2 systems of hyperbolic conservation laws*, J. Amer. Math. Soc., 12 (1999), pp. 729–774.

[24] T.-P. Liu and T. Yang, *Well posedness theory for hyperbolic conservation laws*, Comm. Pure Appl. Math., 52 (1999), pp. 1553–1586.

[25] O. A. Oleinik, *Discontinuous solutions of nonlinear differential equations*, Uspekhi Mat. Nauk, 12 (1957), pp. 3–73 (in Russian); Amer. Math. Soc. Transl. Ser. 2, 26 (1963), pp. 95–172 (in English).

[26] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1982.

[27] B. Temple, *No L^1-contractive metrics for system of conservation laws*, Trans. Amer. Math. Soc., 288 (1985), pp. 471–480.

[28] A. Volpert, *The space BV and quasilinear equations*, Mat. Sb., 73 (1967), pp. 255–302 (in Russian); Mat. USSR-Sb., 2 (1967), pp. 225–267 (in English).

SIAM J. MATH. ANAL.
Vol. 33, No. 2, pp. 440–447

A RESULT ON THE BLOW-UP RATE FOR THE ZAKHAROV SYSTEM IN DIMENSION 3[*]

VINCENT MASSELIN[†]

Abstract. We consider a blow-up solution (u, n, v) of the Zakharov system in $\mathbb{R}^3$:

$$\begin{cases} iu_t = -\Delta u + nu, \\ n_t = -\nabla \cdot v, \\ v_t = -\nabla(n + |u|^2). \end{cases}$$

If T is the finite blow-up time, we show the following integral estimate for n:

$$\int_0^T \left(\int_{\mathbb{R}^3} |n(x,t)|^q dx \right)^{\frac{\gamma}{q}} dt = +\infty,$$

where $\epsilon \in]0, \frac{1}{4}]$, $q = \frac{3}{2(1-\epsilon)} \in \left]\frac{3}{2}, 2\right]$, and $\gamma > \frac{1}{\epsilon}$. In particular, this implies that, for $a < 1$,

$$\sup_{t \in [O,T)} \left((T-t)^{a\epsilon} \left(\int_{\mathbb{R}^3} |n(x,t)|^q dx \right)^{\frac{1}{q}} \right) = +\infty.$$

Key words. Zakharov system, blow-up

AMS subject classifications. 35B40, 35Q99

PII. S0036141099363687

1. Introduction.

In this paper, we consider the three-dimensional (3D) Zakharov system

$$\begin{cases} iu_t = -\Delta u + nu, \\ n_{tt} = \Delta n + \Delta |u|^2, \\ u(0) = u_0, \ n(0) = n_0, \ n_t(0) = n_1, \end{cases} \tag{1.1}$$

where $u : [0, T) \times \mathbb{R}^3 \to \mathbb{C}, n : [0, T) \times \mathbb{R}^3 \to \mathbb{R}$ and u_0, n_0, n_1 are initial data.

In fact, we consider the system (1.1) in the Hamiltonian case. That is, we assume that there is a $w_0 : \mathbb{R}^3 \to \mathbb{R}$ such that

$$n_t(0) = n_1 = -\Delta w_0.$$

Then, for all t, there is a $w(t)$ such that

$$n_t(t) = -\Delta w(t) = -\nabla \cdot v(t),$$

where $v(t) = \nabla w(t)$. In this case, (1.1) can be written in the form

$$\begin{cases} iu_t = -\Delta u + nu, \\ n_t = -\nabla \cdot v, \\ v_t = -\nabla(n + |u|^2), \\ u(0) = u_0, \ n(0) = n_0, \ v(0) = v_0. \end{cases} \tag{1.2}$$

[*] Received by the editors November 3, 1999; accepted for publication (in revised form) November 14, 2000; published electronically July 31, 2001.
http://www.siam.org/journals/sima/33-2/36368.html

[†] Département de Mathématiques, Université de Cergy-Pontoise, 2, avenue Adolphe Chauvin, 95302 Cergy-Pontoise Cédex, France (masselin@math.pst.u-cergy.fr).

The Cauchy problem for (1.1) or (1.2) has been studied in several papers. In [7], Ozawa and Tsutsumi proved that (1.1) is locally (in time) well-posed for initial data $(u_0, n_0, v_0) \in H^2 \times H^1 \times H^1$. With a method introduced by Bourgain for nonlinear dispersive equations, Bourgain and Colliander [1] and Ginibre, Tsutsumi, and Velo [3] have improved this result. The last authors have proved that (1.1) is locally (in time) well-posed for initial data $(u_0, n_0, n_1) \in H^k \times H^l \times H^{l-1}$ provided $l \geq 0$ and $2k - (l+1) \geq 0$; the solution satisfies

$$(u, n, n_t) \in \mathcal{C}\left([0,T), H^k \times H^l \times H^{l-1}\right).$$

Moreover, $\forall t \in [0,T)$,

$$\int_{\mathbb{R}^3} |u(x,t)|^2 dx = \int_{\mathbb{R}^3} |u_0(x)|^2 dx \tag{1.3}$$

and, if $k = 1$,

$$H(t) = H(0), \tag{1.4}$$

where

$$H(t) = \int_{\mathbb{R}^3} |\nabla u|^2 + n|u|^2 + \frac{1}{2}|v|^2 + \frac{1}{2}n^2 \, dx$$

is the Hamiltonian.

Here, we consider a blow-up solution of (1.2) in $\mathcal{C}([0,T), H_1)$, where $H_1 = H^1 \times L^2 \times L^2$: we assume that

$$\begin{cases} T < +\infty, \\ \lim\limits_{t \to T} |(u,n,v)(t)|_{H_1} = +\infty. \end{cases} \tag{1.5}$$

There is no general result for the existence of a blow-up solution but, in [6], Merle proved the following blow-up theorem:

Assume that for all time, $(u,n,v)(t)$ are radially symmetric functions. Moreover, assume that $H(0) < 0$. Then, $(u,n,v)(t)$ blows up. More precisely, we have the following alternatives:

(i) *$(u,n,v)(t)$ blows up in finite time;*

(ii) *$(u,n,v)(t)$ blows up in infinite time in H_1: $(u,n,v)(t)$ is defined for all t and* $\lim_{t \to +\infty} |(u,n,v)(t)|_{H_1} = +\infty$.

We will use the following notations. B is the ball $\{x \in \mathbb{R}^3;\ |x| < 1\}$. For $p \in [1, +\infty]$ and u a function of x or of (x,t), $|u|_p$ will be the L^p-norm in x on $\mathbb{R}^3$. We fix $\epsilon \in (0, \frac{1}{4}]$, $q = \frac{3}{2(1-\epsilon)} \in \left(\frac{3}{2}, 2\right]$ and $p \in [4,6)$ defined by the relation

$$\frac{2}{p} + \frac{1}{q} = 1. \tag{1.6}$$

In particular, if $\epsilon = \frac{1}{4}$, $q = 2$ and $p = 4$. C will represent any constant which depends on ϵ and $|(u_0, n_0, v_0)|_{H_1}$.

In this paper, we prove the following integral estimate on space and time for n.

THEOREM 1.1. *Let $(u,n,v) \in \mathcal{C}([0,T); H_1)$ be a blow-up solution of the Zakharov system* (1.2). *We assume that when $T < \infty$ that (u,n,v) blows up at time T. If $\gamma > \frac{1}{\epsilon}$, then*

$$\int_0^T \left(\int_{\mathbb{R}^3} |n(x,t)|^q dx \right)^{\frac{\gamma}{q}} dt = +\infty. \tag{1.7}$$

With this result, we can easily prove the following estimate.

THEOREM 1.2. *Let $(u,n,v) \in \mathcal{C}([0,T);H_1)$ be a blow-up solution of the Zakharov system* (1.2). *We assume that when $T<\infty$ that (u,n,v) blows up at time T. If $a<1$, then*

$$\sup_{t\in[0,T)} \Big((T-t)^{a\epsilon}|n(t)|_q\Big) = +\infty. \tag{1.8}$$

Assuming the radial symmetry for (u,n,v), we prove an estimate of n in $L^q(B)$ as follows.

THEOREM 1.3. *Let us assume that $(u,n,v) \in \mathcal{C}([0,T);H_1)$ is a radially symmetric solution of* (1.2). *We assume that when $T<\infty$ that (u,n,v) blows up at time T. If $a \in (0,\frac{1}{3})$, then*

$$\sup_{t\in[0,T)} \Big((T-t)^{a\epsilon}\,|n(t)|_{L^q(B)}\Big) = +\infty. \tag{1.9}$$

In [4], Landman et al. have worked out a numerical computation which suggests there is solution of (1.2) which blows up with the profile

$$\begin{cases} \tilde{u}(x,t) = \dfrac{2}{3(T-t)}P\Big(\dfrac{x}{\sqrt{3}(T-t)^{\frac{2}{3}}}\Big)e^{i(T-t)^{-\frac{1}{3}}}, \\ \tilde{n}(x,t) = \dfrac{1}{3(T-t)^{\frac{4}{3}}}N\Big(\dfrac{x}{\sqrt{3}(T-t)^{\frac{2}{3}}}\Big), \\ \tilde{v}(x,t) = \dfrac{2}{3}\dfrac{1}{(T-t)^{\frac{5}{3}}}w\Big(\dfrac{x}{\sqrt{3}(T-t)^{\frac{2}{3}}}\Big)\dfrac{x}{r}, \end{cases} \tag{1.10}$$

where (P,N,w) are radially symmetric and the solution of the system

$$\begin{cases} \Delta P = P + NP, \\ 5w + 2rw_r = -(P^2)_r, \\ 4w + 2rw_r = -(r^2N)_r \end{cases} \tag{1.11}$$

with $r=|x|$. This system is equivalent to

$$\begin{cases} \Delta P = P + NP, \\ \frac{1}{2}(2r^2N_{rr} + 13rN_r + 14N) = \Delta P^2. \end{cases} \tag{1.12}$$

In [5], we prove that there exist infinitely many radial and $\mathcal{C}^\infty$ solutions of (1.11), (1.12) such that P is positive decreasing and satisfying $\lim_{r\to+\infty} P(r) = 0$. Moreover, for such a solution

$$\lim_{r\to+\infty} 3r^2N(r) = -2P^2(0) \neq 0.$$

Then, when t tends to T, $|\tilde{n}(t)|_{L^q(B)}$ is equivalent to $(T-t)^{-\frac{4\epsilon}{3}}|N|_{L^q(\mathbb{R}^3)}$, and for all $\gamma > \frac{3}{4\epsilon}$

$$\int_0^T \Big(\int_{\mathbb{R}^3} |\tilde{n}(x,t)|^q dx\Big)^{\frac{\gamma}{q}} dt = +\infty.$$

Therefore, the result of Theorem 1.1 is not exactly optimal: we prove this for $\gamma > \frac{1}{\epsilon}$ instead of $\gamma > \frac{3}{4\epsilon}$. Moreover, we don't know if there exists $c>0$ such that $|n(t)|_{L^q(B)} \geq c(T-t)^{\frac{4}{3}\epsilon}$. We can prove only the following:

$$\lim_{t\to T} |n(t)|_{L^q(B)} = +\infty.$$

(See the end of the proof of Theorem 1.3.)

In the second part of this paper we prove the results. First, we recall some useful results. Then, we use the conservation of the Hamiltonian, the dispersion effect of the Schrödinger group $e^{it\Delta}$, and a Gronwall lemma to prove the space-time estimate (see Theorem 1.1). Finally, we prove Theorem 1.3.

2. Proofs of the theorems.

2.1. Some general results. First, let us state some inequalities with our notations, as follows.

LEMMA 2.1.

- *If* $\phi \in H^1(\mathbb{R}^3)$ *is radially symmetric, then* $\phi \in L^\infty(\mathbb{R}^3 \setminus B)$ *and*

$$|\phi|^2_{L^\infty(\mathbb{R}^3\setminus B)} \leq C|\nabla\phi|_2|\phi|_2. \tag{2.1}$$

- *If* Ω *is a domain in* $\mathbb{R}^3$ *and* $\phi \in H^1(\Omega)$, *then* $\phi \in L^p(\Omega)$ *and*

$$|\phi|_{L^p(\Omega)} \leq C|\nabla\phi|^{1-\epsilon}_{L^2(\Omega)}|\phi|^{\epsilon}_{L^2(\Omega)}, \tag{2.2}$$

in particular, $\phi \in L^4(\Omega)$ *and*

$$|\phi|_{L^4(\Omega)} \leq C|\nabla\phi|^{\frac{3}{4}}_{L^2(\Omega)}|\phi|^{\frac{1}{4}}_{L^2(\Omega}. \tag{2.3}$$

- *Let* $S(t) = e^{it\Delta}$ *and* $p' = \frac{p}{p-1}$. *There exists a constant* $C > 0$ *such that* $\forall\phi \in L^{p'}(\mathbb{R}^3)$ *and* $\forall t > 0$,

$$|S(t)\phi|_p \leq \frac{C}{t^{1-\epsilon}}|\phi|_{p'}. \tag{2.4}$$

For a blow-up solution, we have the following limits, which are more precise than (1.5).

LEMMA 2.2. *If* (u, n, v) *is a blow-up solution in* H_1 *and* T *is the finite blow-up time, then*

$$\lim_{t\to T}|u(t)|_4 = +\infty, \tag{2.5}$$

$$\lim_{t\to T}|\nabla u(t)|_2 = +\infty. \tag{2.6}$$

Proof. If (2.5) is false, then there is a sequence (t_k) and a constant c such that $\lim t_k = T$ and $\forall\ k$

$$|u(t_k)|_4 \leq c.$$

Then, according to the conservation of the Hamiltonian and Hölder inequalities,

$$\begin{aligned}\frac{1}{2}\int_{\mathbb{R}^3} n^2(x,t_k)dx &\leq H + \left|\int_{\mathbb{R}^3} n(x,t_k)|u|^2(x,t_k)dx\right| \\ &\leq H + |n(t_k)|_2|u(t_k)|_4^2.\end{aligned}$$

Therefore, $(|n(t_k)|_2)$ is bounded. Then, using again the conservation of H, we get

$$\int_{\mathbb{R}^3}|\nabla u(x,t_k)|^2dx + \frac{1}{2}\int_{\mathbb{R}^3} n^2(x,t_k) + |v(x,t_k)|^2dx \leq H + |n(t_k)|_2|u(t_k)|_4^2 \leq C,$$

which contradicts the blow-up assumption. □

Now (2.6) comes from the inequality (2.3) and the conservation of the L^2-norm.

We prove an integral inequality on $|u(t)|_p$ and $|n(t)|_q$ as follows.

LEMMA 2.3. *There exists a positive constant C such that $\forall t \in [0, T)$,*

$$|u(t)|_p \leq C\left(1 + \int_0^t \frac{1}{(t-s)^{1-\epsilon}} |n(s)|_q |u(s)|_p ds\right) \tag{2.7}$$

and, in particular,

$$|u(t)|_4 \leq C\left(1 + \int_0^t \frac{1}{(t-s)^{\frac{3}{4}}} |n(s)|_2 |u(s)|_4 ds\right). \tag{2.8}$$

Proof. We write the equation $iu_t = -\Delta u + nu$ on the integral form:

$$u(t) = S(t)u_0 - i\int_0^t S(t-s)n(s)u(s)ds.$$

Then, according to the Minkowski inequality,

$$|u(t)|_p \leq |S(t)u_0|_p + \int_0^t |S(t-s)n(s)u(s)|_p ds.$$

On the one hand, by (2.4), we have

$$|S(t-s)n(s)u(s)|_p \leq \frac{c}{(t-s)^{1-\epsilon}} |n(s)u(s)|_{p'}.$$

But, $1/p + 1/q = 1 - 1/p = 1/p'$, so according to the Hölder inequality,

$$|n(s)u(s)|_{p'} \leq |n(s)|_q |u(s)|_p$$

and

$$|S(t-s)n(s)u(s)|_p \leq \frac{1}{(t-s)^{1-\epsilon}} |n(s)|_q |u(s)|_p.$$

On the other hand, according to (2.2),

$$\begin{aligned} |S(t)u_0|_p &\leq c|S(t)u_0|_2^{\epsilon} |\nabla(S(t)u_0)|_2^{1-\epsilon} \\ &\leq c|u_0|_2^{\epsilon} |\nabla u_0|_2^{1-\epsilon} \end{aligned}$$

and (2.7) comes. In the particular case $\epsilon = 1/4,\ p = 4,\ q = 2$, we obtain (2.8). □

2.2. Proof of the integral estimate. By contradiction, let us assume there exists $\gamma_0 > \frac{1}{\epsilon}$ such that

$$\int_0^T \left(\int_{\mathbb{R}^3} |n(x,t)|^q dx\right)^{\frac{\gamma_0}{q}} dt < +\infty. \tag{2.9}$$

By Lemma 2.3, we have

$$|u(t)|_p \leq C\left(1 + \int_0^t \frac{1}{(t-s)^{1-\epsilon}} |n(s)|_q |u(s)|_p ds\right),$$

where $1-\epsilon+\frac{1}{\gamma}<1$ and $|n(s)|_q \in L^{\gamma}(0,T)$. Therefore, a Gronwall lemma (see, for example, [2, Lemma 8.1.1]) implies that there exists a constant $c>0$ such that $\forall\, t \in [0,T)$,

$$|u(t)|_p \leq C. \tag{2.10}$$

Now we display two cases. At first, if $\epsilon = \frac{1}{4}$, then $p=4$ and (2.10) contradicts Lemma 2.2. Also, we have shown that $\forall \gamma > 4$,

$$\int_0^T \left(\int_{\mathbb{R}^3} |n(x,t)|^2 dx \right)^{\frac{\gamma}{2}} dt = +\infty. \tag{2.11}$$

Now we return to the general case. By the conservation of the Hamiltonian,

$$\frac{1}{2}|n(t)|_2^2 \leq H + \left| \int_{\mathbb{R}^3} n(x,t)|u(x,t)|^2 dx \right|.$$

But $\frac{1}{q}+\frac{1}{p/2}=1$, so, by the Hölder inequality,

$$\begin{aligned} \frac{1}{2}|n(t)|_2^2 &\leq H + |n(t)|_q |u(t)|_p^2 \\ &\leq H + C|n(t)|_q. \end{aligned}$$

Therefore,

$$\int_0^T \left(\int_{\mathbb{R}^3} |n(x,t)|^2 dx \right)^{\frac{2\gamma_0}{2}} dt < +\infty$$

with $2\gamma_0 > \frac{2}{\epsilon} \geq 8$. So this contradicts the previous case $\epsilon = \frac{1}{4}$ and concludes the proof of the space-time estimate.

2.3. The radial case. We consider $a \in (0, \frac{1}{3})$ and we assume that (u,n,v) is a radially symmetric blow-up solution such that $\forall\, t$, $(u,n,v)(t) \in H^1 \times L^2 \times L^2$ and

$$|n(t)|_{L^q(B)} \leq \frac{c}{(T-t)^{a\epsilon}}. \tag{2.12}$$

First, we show the following estimate of $n(t)$ in L^2:

$$|n(t)|_2 \leq \frac{C}{(T-t)^{\frac{3a}{4}}}. \tag{2.13}$$

According to the conservation of the Hamiltonian, we have

$$\begin{aligned} \int_{\mathbb{R}^3} |\nabla u(x,t)|^2 dx + \frac{1}{2}\int_{\mathbb{R}^3} n(x,t)^2 dx \leq H &+ \left| \int_B n(x,t)|u(x,t)|^2 dx \right| \\ &+ \left| \int_{\mathbb{R}^3 \setminus B} n(x,t)|u(x,t)|^2 dx \right|, \end{aligned}$$

but $\frac{1}{q}+\frac{1}{p/2}=1$, so by the Hölder inequality, (2.2), and (2.12),

$$\begin{aligned} \left| \int_B n(x,t)|u(x,t)|^2 dx \right| &\leq \left(\int_B |n(x,t)|^q dx \right)^{\frac{1}{q}} \left(\int_B |u(x,t)|^p dx \right)^{\frac{2}{p}} \\ &\leq C \frac{1}{(T-t)^{a\epsilon}} \left| \nabla u(t) \right|_2^{2(1-\epsilon)}. \end{aligned}$$

On the other hand,

$$\begin{aligned}\left|\int_{\mathbb{R}^3\setminus B} n(x,t)|u(x,t)|^2dx\right| &\leq \frac{1}{2}\int_{\mathbb{R}^3\setminus B} n(x,t)^2dx + \frac{1}{2}\int_{\mathbb{R}^3\setminus B} |u(x,t)|^4dx\\ &\leq \frac{1}{2}\int_{\mathbb{R}^3\setminus B} n(x,t)^2dx + \frac{1}{2}|u(t)|^2_{L^\infty(\mathbb{R}^3\setminus B)}|u(t)|_2^2\\ &\leq \frac{1}{2}\int_{\mathbb{R}^3\setminus B} n(x,t)^2dx + C|\nabla u(t)|_2\end{aligned}$$

by the inequality (2.1). Then

$$\left(\int_{\mathbb{R}^3} |\nabla u(x,t)|^2dx\right)^{\epsilon} \leq H + \left(\int_{\mathbb{R}^3} |\nabla u(x,t)|^2dx\right)^{\epsilon-1} + \frac{c}{(T-t)^{a\epsilon}}.$$

However, by Lemma 2.2 (see (2.6)), $\lim \left(\int |\nabla u|^2\right)^{\epsilon-1} = 0$, so

$$\int_{\mathbb{R}^3} |\nabla u(x,t)|^2dx \leq \frac{c}{(T-t)^{\frac{a}{2}}}.$$

Then, by (2.3), we get

$$|u(t)|_4 \leq \frac{C}{(T-t)^{\frac{3a}{8}}}.$$

By the conservation of the Hamiltonian,

$$\frac{1}{2}\int_{\mathbb{R}^3} n^2(x,t)dx \leq H + |n(t)|_2|u(t)|_4^2$$

and (2.13) follows. Then, for $\gamma \in]4, \frac{4}{3a}[$ $(a < \frac{1}{3})$,

$$\int_0^T |n(t)|_2^\gamma dt < +\infty,$$

which contradicts Theorem 1.2 and concludes the proof. To show that $|n(t)|_{L^q(B)} \to +\infty$ as $t \to T$, we use the same method: if there exists $t_k \to T$ such that $|n(t_k)|_{L^q(B)} \leq c$, then we prove as above that $\int_{\mathbb{R}^3} |\nabla u(x,t_k)|^2dx \leq c$.

Acknowledgment. The author is grateful to Frank Merle for his direction and help during the preparation of this work.

REFERENCES

[1] J. Bourgain and J. Colliander, *On wellposedness of the Zakharov system*, Internat. Math. Res. Notices, 11 (1996), pp. 515–546.

[2] T. Cazenave and A. Haraux, *Introduction aux problèmes d'évolution semi-linéaires*, Math. Appl. 1, Ellipses, Paris, 1990.

[3] J. Ginibre, Y. Tsutsumi, and G. Velo, *On the Cauchy problem for the Zakharov system*, J. Funct. Anal., 151 (1997), pp. 384–436.

[4] M. Landman, G.C. Papanicolaou, C. Sulem, P.L. Sulem, and X.P. Wang, *Stability of isotropic self-similar dynamics for scalar collapse*, Phys. Rev. A (3), 46 (1992), pp. 7869–7876.

[5] V. Masselin, *Existence of a Solution for a System Related to the Singularity for* 3*D Zakharov System*, preprint.

[6] F. Merle, *Blow-up results of viriel type for Zakharov equations*, Comm. Math. Phys., 175 (1996), pp. 433–455.

[7] T. Ozawa and Y. Tsutsumi, *Existence and smoothing effect of solutions for the Zakharov equations*, Publ. Res. Inst. Math. Sci., 28 (1992), pp. 329–361.

SIAM J. MATH. ANAL.
Vol. 33, No. 2, pp. 448–460

THE STABILITY OF SUBDIVISION OPERATOR AT ITS FIXED POINT*

VLADIMIR PROTASOV†

Abstract. We consider the univariate two-scale refinement equation $\varphi(x) = \Sigma_{k=0}^{N} c_k \varphi(2x-k)$, where $c_0, \dots, c_N$ are complex values and $\Sigma c_k = 2$.

This paper analyzes the correlation between the existence of smooth compactly supported solutions of this equation and the convergence of the corresponding cascade algorithm/subdivision scheme. We introduce a criterion that expresses this correlation in terms of the mask of the equation. We show that the convergence of the subdivision scheme depends on values that the mask takes at the points of its *generalized cycles.* This means in particular that the stability of shifts of refinable function is not necessary for the convergence of the subdivision process. This also leads to some results on the degree of convergence of subdivision processes and on factorizations of refinable functions.

Key words. refinement equations, cascade algorithm, subdivision process, degree of convergence, stability, cycles, tree

AMS subject classifications. 26C10, 39B32, 42A05, 42A38

PII. S0036141099356283

1. Introduction.

Refinement equations have been studied by many authors in great detail in connection with their role in the study of wavelets and of subdivision schemes in approximation theory and the design of curves and surfaces. In this paper we study the correlation between the existence of smooth solutions of refinement equations and the convergence of the corresponding subdivision schemes. We restrict ourselves to univariate equations having compactly supported mask. We obtain a criterion for the convergence of subdivision process under the condition that the associated refinement equation has a smooth solution.

Throughout the paper we denote by $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$ the unit circle, by $\mathcal{H}$ the space of entire functions on $\mathbb{C}$, by $\mathcal{C}^l$ the space of l times continuously differentiable functions on $\mathbb{R}$, by $\mathcal{C}^0 = \mathcal{C}$ the space of continuous functions, by $\mathcal{C}_0^l$ the space of compactly supported functions from $\mathcal{C}^l$, and by $\mathcal{C}_0$ the space of compactly supported continuous functions on $\mathbb{R}$. A sequence $\{f_k\}$ converges to zero in $\mathcal{C}_0^l$ if it converges to zero in $\mathcal{C}^l$ and the supports of $f_k,\ k \in \mathbb{N}$ are uniformly bounded.

Consider a refinement equation

$$\varphi(x) = \sum_{k=0}^{N} c_k \varphi(2x-k), \tag{1.1}$$

where $c_k \in \mathbb{C}$, $\sum_k c_k = 2$. It is well known that a $\mathcal{C}_0$-solution of this equation (*refinable function*), if it exists at all, is unique up to normalization, has its support on the

[*] Received by the editors May 17, 1999; accepted for publication December 12, 2000; published electronically July 31, 2001.

http://www.siam.org/journals/sima/33-2/35628.html

[†] Department of Mechanics and Mathematics, Moscow State University, Vorobyovy Gory 3, Moscow, 119899, Russia (protasov@dionis.iasnet.ru). Current address: Erasmus Universiteit Rotterdam, Fac. der Economische Wetenschappen, H 11-21, Postbus 1738, 3000 DR Rotterdam, The Netherlands (protassov@few.eur.nl).

segment $[0, N]$, and can be represented in the frequency domain by the formula

$$\widehat{\varphi}(\xi) = \widehat{\varphi}(0) \prod_{r=1}^{\infty} m\left(\frac{\xi}{2^r}\right), \tag{1.2}$$

where $m(\xi) = \frac{1}{2}\sum_{k=0}^{N} c_k e^{-ik\xi}$ is the *mask* of (1.1) (as usually we denote $\hat{f}(\xi) = \int f(x)e^{-i\xi x}dx$). For a given mask $a(\xi)$ let us denote by $[a]$ the corresponding refinement equation. Let us also define the subspaces of the space $\mathcal{C}_0$ as

$$\mathcal{M}^l = \{f \in \mathcal{C}_0 \mid \widehat{f}(\xi)(1 - e^{-i\xi})^{-l-1} \in \mathcal{H}\}, \quad l \geq 0, \tag{1.3}$$

and the subspaces of $\mathcal{C}_0^l$ as

$$\mathcal{L}^l = \{f \in \mathcal{C}_0^l \mid \widehat{f^{(l)}} \in \mathcal{M}^l\}, \quad l \geq 0.$$

In other words the Fourier transform of a function from M^l has zeros of order $\geq l+1$ at all the points $2\pi k,\ k \in \mathbb{Z}$. The Fourier transform of a function from $\mathcal{L}^l$ has zero at the point $\xi = 0$ and has zeros of order $\geq l+1$ at all the points $2\pi k,\ k \in \mathbb{Z} \setminus \{0\}$.

Let us also denote $\mathcal{L} = \mathcal{L}^0 = \mathcal{M}^0$. By Poisson summation formula we have

$$f \in \mathcal{L} \ \Leftrightarrow\ f \in \mathcal{C}_0,\ \sum_k f(x-k) \equiv 0.$$

The *cascade algorithm* for refinement equations was introduced in [D]. A single iteration of that algorithm is $f_n = Tf_{n-1}$, where f_0 is an initial function from $\mathcal{C}_0$ and $Tf(x) = \sum_k c_k f(2x-k)$ is the *subdivision operator* associated to (1.1). This operator is defined on the space $\mathcal{C}_0$ and has the form

$$\widehat{Tf}(\xi) = m(\xi/2)\widehat{f}(\xi/2) \tag{1.4}$$

in the frequency domain. If f_n converges in the space $\mathcal{C}^l$ to a function $\varphi \in \mathcal{C}_0^l$ $(l \geq 0)$, then obviously it converges in $\mathcal{C}_0^l$ and φ is the solution of (1.1). Moreover, in that case the function $g = f_0 - \varphi$ necessarily belongs to $\mathcal{L}^l$ (see [CDM], [Du1]). The cascade algorithm *converges in* $\mathcal{C}^l$ if $T^n g \to 0,\ n \to \infty$ for any $g \in \mathcal{L}^l$. Properties of the cascade algorithms have been studied by many authors in various contexts. This algorithm gives a simple way for approximation of refinable functions. In particular, this was put to good use in the study of wavelets [D],[DL1], [Du2]. On the other hand, the convergence of the cascade algorithm is equivalent to the convergence of the corresponding *subdivision scheme* (see [RS] for many references). For a given mask $m(\xi)$, we say that the *subdivision process* $\{m\}$ *converges in* $\mathcal{C}^l$ if the corresponding cascade algorithm or the corresponding subdivision scheme converges in that space.

It is clear that the convergence of the subdivision process in $\mathcal{C}^l$ implies that the corresponding refinement equation has a $\mathcal{C}_0^l$-solution. In general, the converse is not true. (See [DL2] and [CDM] for many examples. See also [CH], [W], [RS] for general discussions of this aspect.) In this paper we analyze the correlation between the existence of smooth solutions of refinement equations and the convergence of the corresponding subdivision process. In other words, we study stability of the subdivision operator at its fixed point. Let us first formulate several previously known results on this problem.

2. Preliminary results. Necessary conditions for the convergence of subdivision processes were first introduced in the work [DGL2].

If a subdivision process $\{m\}$ converges in $\mathcal{C}^l$, then its mask can be factored as

$$m(\xi) = \left(\frac{1+e^{-i\xi}}{2}\right)^{l+1} a(\xi) \tag{2.1}$$

for some trigonometric polynomial $a(\xi)$. In particular, the condition

$$m(\xi) = \left(\frac{1+e^{-i\xi}}{2}\right) a(\xi) \tag{2.2}$$

is necessary for the convergence of the subdivision process in $\mathcal{C}$ [DGL2].

For a given mask m denote by $\mathbf{l}(m)$ the maximal integer l such that condition (2.1) is satisfied. So if a subdivision process $\{m\}$ converges in $\mathcal{C}^k$, then $k \le \mathbf{l}(m)$. Let us remark that condition (2.1) is not necessary for the existence of $\mathcal{C}_0^l$-solutions of the refinement equation [DL2], [P2].

Sufficient conditions for the convergence of the subdivision process in the space $\mathcal{C}$ (i.e., in the case $l = 0$) were introduced in [CDM].

If a refinement equation $[m]$ has a $\mathcal{C}_0$-solution and that solution is stable in the space $L^\infty(\mathbb{R})$ (i.e., its integer translates possess Riesz basis property in that space), then the subdivision process $\{m\}$ converges in $\mathcal{C}$ [CDM].

This condition is simplified by the criterion of stability of refinable functions proved in [JW] and [Z] and introduced independently in [He1]. To formulate it we need some notation. Let $p(\xi)$ be a trigonometric polynomial. If for some $\alpha \in \mathbb{T}$ we have $p(\alpha/2) = p(\pi + \alpha/2) = 0$, then the pair $\{\alpha/2, \pi + \alpha/2\}$ is a *pair of symmetric roots* for $p(\xi)$. In order to be defined, we set that for any $\alpha \in \mathbb{T}$ the value $\alpha/2 \in \mathbb{T}$ has the corresponding real value from the half-interval $[0, \pi)$. Further, a given set $\mathbf{b} = \{\beta_1, \dots, \beta_n\} \subset \mathbb{T}$, where $n \ge 2$, is called a *cycle* of the polynomial $p(\xi)$ if $2\beta_j = \beta_{j+1}$ for $j = 1, \dots, n$ (we set $\beta_{n+1} = \beta_1$) and $p(\beta_j + \pi) = 0$ for all $j = 1, \dots, n$. We consider only irreducible cycles, i.e., we suppose everywhere that all elements of a cycle are different. Now let us remember the criterion of stability of refinable functions.

The $\mathcal{C}_0$-solution of a refinement equation is stable in L^∞ if and only if its mask has neither symmetric roots nor cycles [JW], [Z], [He1].

Those two results can be summarized in the following theorem.

Theorem 2.1 (see [CDM],[JW],[Z],[He1]). *Suppose a mask m satisfying* (2.2) *has neither symmetric roots nor cycles; if the equation $[m]$ has a $\mathcal{C}_0$-solution, then the process $\{m\}$ converges in $\mathcal{C}$.*

Remark 1. The statement of Theorem 2.1 can also be formulated in terms of *Cohen's criterion* (see [D]). Namely, it was shown in [V, Proposition 2.4] that a mask satisfies Cohen's criterion if and only if it has neither symmetric roots nor cycles.

3. Statement of the fundamental theorems. In this paper we give a criterion of stability of subdivision operator at its fixed point (Theorem 3.1). We will see that symmetric roots of mask do not influence the convergence of subdivision process (Corollary 3). This means in particular that the stability of solutions is not necessary for the convergence of the subdivision process. The convergence depends on values of the mask at the points of cycles.

To formulate the criterion we need some further notation. Everywhere below we consider trigonometric polynomials without positive powers, i.e., polynomials of the form $p(\xi) = \sum_{k=0}^{N} a_k e^{-ik\xi}$. As usual we set deg $p = N$ (assuming $a_N \neq 0$).

To an arbitrary trigonometric polynomial p we associate a polynomial $R[p]$ as follows: suppose $r(\xi)$ is the polynomial of smallest degree such that the function $\frac{p(\xi)r(\xi)}{r(2\xi)}$ is a polynomial without symmetric roots; then we set $R[p](\xi) = \frac{p(\xi)r(\xi)}{r(2\xi)}$. The reader will have no difficulty in showing that the mapping $p \mapsto R[p]$ is well defined. For given p, the polynomial $R[p]$ can by easily found algorithmically. If p has no symmetric roots, then $R[p] = p$. If $\{\alpha/2, \pi + \alpha/2\}$ is a pair of symmetric roots of p, then we pass from $p(\xi)$ to the polynomial $p_\alpha(\xi) = \frac{p(\xi)(1-e^{i(\alpha-\xi)})}{1-e^{i(\alpha-2\xi)}}$. After several steps we obtain a polynomial $\tilde{p}(\xi)$ that has no symmetric roots. In general, there exist several different ways to realize each step of this algorithm: if there exist several pairs of symmetric roots, we can choose any of them to pass to the next polynomial. Nevertheless, *the result (i.e., the polynomial $\tilde{p}(\xi)$) does not depend on that choice and coincides with the polynomial $R[p]$*. The proof of this fact is left to the reader.

For any trigonometric polynomial p and any finite subset $Y = \{\alpha_1, \dots, \alpha_n\} \subset \mathbb{T}$, we denote $\rho_p(Y) = (\prod_{q=1}^{n} |p(\alpha_q)|)^{1/n}$. If the set Y is cyclic (i.e, $\alpha_{q+1} = 2\alpha_q$, $q = 1, \dots, n$, where $\alpha_{n+1} = \alpha_1$), then $\rho_p(Y) = \rho_{R[p]}(Y)$. (The proof is trivial.)

Now let us formulate the criterion of stability of subdivision process.

THEOREM 3.1. *Suppose a refinement equation $[m]$ has a $\mathcal{C}_0^l$-solution $l \geq 0$; then the process $\{m\}$ converges in $\mathcal{C}^l$ if and only if the mask m satisfies* (2.1) *and for any cycle $\mathbf{b}$ of the polynomial $R[m]$ we have $\rho_m(\mathbf{b}) < 2^{-l}$.*

The simplest corollary of this theorem is the following generalization of Theorem 2.1 from the case $l = 0$ to an arbitrary integer factor $l \geq 0$.

COROLLARY 1. *Suppose a mask m satisfying* (2.1) *has neither symmetric roots nor cycles; if the equation $[m]$ has a $\mathcal{C}_0^l$-solution, then the process $\{m\}$ converges in $\mathcal{C}^l$.*

Another problem is to explore the degree of convergence of the subdivision processes. For a given integer $l \geq 0$, a mask m, and a function $f \in \mathcal{L}^l$ denote

$$\nu_l(m, f) = -\lim_{n\to\infty} \frac{\log_2 \|T^n[f^{(l)}]\|_{\mathcal{C}}}{n},$$

where T is the subdivision operator associated to m. (We set $\log_2 0 = -\infty$.) Also for a subspace $\mathcal{V} \subset \mathcal{L}^l$ we denote $\nu_l(m, \mathcal{V}) = \inf_{f\in\mathcal{V}} \nu_l(m, f)$. The value $\nu_l(m) = \nu_l(m, \mathcal{L}^l)$ is the *degree of convergence of the process $\{m\}$ in the space $\mathcal{C}^l$*.

For any mask m we have $\nu_l(m) \leq l + 1$ (see [DL1]). Furthermore, it was shown in [DL1] and [HC] that a process $\{m\}$ converges in $\mathcal{C}^l$ if and only if $\nu_l(m) > l$. In particular, the inequality $\nu_0(m) > 0$ means that $\{m\}$ converges in $\mathcal{C}$. Let L be the maximal integer such that $\{m\}$ converges in $\mathcal{C}^L$. (If the process $\{m\}$ does not converge in $\mathcal{C}$, then we set $L = 0$.) The values $\nu_l(m)$, $l = 0, 1, \dots$ are connected as follows:

$$\nu_l(m) = l + 1 \quad \text{for} \quad l < L; \quad \nu_l(m) = \nu_L(m) \quad \text{for} \quad l \geq L. \tag{3.1}$$

The proof can be found in [DL2]. The value $\nu_L(m)$ is said to be the *degree of convergence of the process* $\{m\}$ and is denoted in what follows by $\nu(m)$. Thus, if $\nu(m_1) = \nu(m_2)$, then $\nu_l(m_1) = \nu_l(m_2)$ for any $l \geq 0$.

The degree of convergence of subdivision processes in various functional spaces was studied in [CDM], [W], [Du1], [Du2], [R3], [RS]. The following theorem reduces this problem (in the space $\mathcal{C}^l$) from general refinement equations to the case of refinement equations having stable solutions.

THEOREM 3.2. *For a given mask m satisfying* (2.1) *for some integer $l \geq 0$ denote $m_1(\xi) = R[m](\xi)/\prod_{k=1}^{q}\prod_{\beta\in\mathbf{b}_k}(1 + e^{i(\beta-\xi)})$, where $\{\mathbf{b}_1, \dots, \mathbf{b}_q\}$ is the set of cycles* *of the polynomial $R[m]$ (counting with multiplicity). Then we have the following: the equation $[m]$ has a $\mathcal{C}_0^l$-solution if and only if $[m_1]$ does; furthermore,*

$$\nu_l(m) = \min\{\nu_l(m_1), -\log_2 \rho_m(\mathbf{b}_1), \ldots, -\log_2 \rho_m(\mathbf{b}_q)\}.$$

COROLLARY 2. *Under the conditions of Theorem* 3.2 *we have*

$$\nu_k(m) = \min\{\nu_k(m_1), -\log_2 \rho_m(\mathbf{b}_1), \ldots, -\log_2 \rho_m(\mathbf{b}_q)\} \qquad \text{for any } k \le l.$$

Moreover, if $\mathbf{l}(m) = \mathbf{l}(m_1)$, *then*

$$\nu(m) = \min\{\nu(m_1), -\log_2 \rho_m(\mathbf{b}_1), \ldots, -\log_2 \rho_m(\mathbf{b}_q)\}.$$

Remark 2. Since the mask m_1 has neither symmetric roots nor cycles, it follows that the $\mathcal{C}_0^l$-solution of the equation $[m_1]$ is stable. Some previously known results on subdivision processes deal with the stable case (see, for instance, [CDM]). Theorem 3.2 makes it possible to extend those results to the case of general refinement equations.

COROLLARY 3. *For an arbitrary mask m satisfying* (2.1) *we have*

$$\nu_l(m) = \nu_l(R[m]).$$

Moreover, in the case $\mathbf{l}(m) = \mathbf{l}(R[m])$ *we have* $\nu(m) = \nu(R[m])$.

To prove this, it is sufficient to apply Theorem 3.2 to the masks m and $R[m]$ and note that $\rho_m(\mathbf{b}_i) = \rho_{R[m]}(\mathbf{b}_i)$.

Thus symmetric roots of the mask do not have influence on the degree of convergence of the subdivision process. So the sufficient conditions from Corollary 1 are not necessary for the convergence.

Remark 3. It can easily be shown that $\mathbf{l}(m) \le \mathbf{l}(R[m])$ for any mask m. There are masks such that $\mathbf{l}(m) < \mathbf{l}(R[m])$ and, moreover, $\nu(m) < \nu(R[m])$. That is why the condition $\mathbf{l}(m) = \mathbf{l}(R[m])$ is essential in the statement of Corollary 3 (see [P2]).

Remark 4. (The degree of convergence in various subspaces of $\mathcal{C}_0$). Consider the family of embedded subspaces $\{\mathcal{M}^l\}$ defined from (1.3). It was shown in [DL2],[Du1] that $f \in \mathcal{M}^l$ whenever $\nu_0(m, f) > l$. So the subspaces $\{\mathcal{M}^l\}$ can be considered as spaces of fast convergence of the subdivision processes. Moreover, if $\nu_0(m, \mathcal{M}^l) > l$, then the mask m satisfies (2.1) and hence all the subspaces $\mathcal{M}^k,\ k = 0, \cdots, l$, are invariant with respect to the corresponding subdivision operator. So it is natural to restrict a subdivision operator to suitable subspace $\mathcal{M}^l$ and consider the value $\nu_0(m, \mathcal{M}^l)$ instead of $\nu_l(m)$ (see, for instance, [CDM], [Du1], [Du2]). Theorems 3.1 and 3.2 of this paper can be reformulated in those terms without any change.

Theorems 3.1 and 3.2 will be proved in the next section. Then, in section 5, we introduce the notion of *generalized cycles* and establish a correlation between zeros of mask m and cycles of the polynomial $R[m]$. As a corollary we shall formulate the criterion of Theorem 3.1 in terms of zeros of the mask m (without the transfer to the polynomial $R[m]$).

4. Proof of the main results. To prove Theorems 3.1 and 3.2 let us first consider the case $l = 0$. The proof will be split into several lemmas and propositions.

For a finite family of real values $\Delta = \{\delta_1, \ldots, \delta_n\}$ (that may coincide), let

$$\mathcal{C}_0\{\Delta\} = \mathcal{C}_0\{\delta_1, \ldots, \delta_n\} = \left\{ f \in \mathcal{C}_0 \mid \widehat{f}(\xi) / \prod_{q=1}^{n} (1 - e^{i(\delta_q - \xi)}) \in \mathcal{H} \right\}.$$

It is clear that $\mathcal{M}^l = \mathcal{C}_0\{0, \cdots, 0\}$ ($l+1$ zeros). From the Poisson summation formula it follows that for any $f \in \mathcal{C}_0\{\Delta\}$ we have

$$\sum_{k\in\mathbb{Z}} e^{ik\delta_q} f(x-k) = 0, \quad q = 1, \dots, n. \tag{4.1}$$

Let us also denote

$$\mathcal{L}_\Delta = \mathcal{C}_0\{0, \Delta\} = \mathcal{C}_0\{0, \delta_1, \dots, \delta_n\} \text{ and } \mathcal{L}_\Delta[0, N] = \{f \in \mathcal{L}_\Delta \mid \operatorname{supp} f \in [0, N]\}.$$

For given $\delta \in \mathbb{R}$, consider the difference operator S_δ acting from the space $\mathcal{C}_0\{\Delta\}$ into the space $\mathcal{C}_0\{\Delta, \delta\} = \mathcal{C}_0\{\delta_1, \dots, \delta_n, \delta\}$ and defined by the formula $S_\delta \psi(x) = \psi(x) - e^{i\delta}\psi(x-1)$.

Lemma 4.1. *For any $\delta \in \mathbb{R}$ the operator S_δ is a homeomorphism of the spaces $\mathcal{C}_0\{\Delta\}$ and $\mathcal{C}_0\{\Delta, \delta\}$.*

Proof. For arbitrary $\varphi \in \mathcal{C}_0\{\Delta, \delta\}$ denote $\psi(x) = S_\delta^{-1}\varphi(x) = \sum_{k=0}^{+\infty} e^{ik\delta}\varphi(x-k)$. If $\operatorname{supp}\varphi \subset [a, b]$ for some integers a, b, then by (4.1) we have $\operatorname{supp}\psi \subset [a, b-1]$. Thus, $\psi \in \mathcal{C}_0$. It now follows that $\psi \in \mathcal{C}_0\{\Delta\}$. It remains to note that $S_\delta\psi = \varphi$ and the operators S_δ and S_δ^{-1} are obviously continuous. □

The following proposition is the first step in the proof of Theorems 3.1 and 3.2.

Proposition 1. *Suppose a mask $m(\xi)$ satisfying* (2.2) *possesses a pair of symmetric roots $\alpha/2$ and $\pi + \alpha/2$. Let $m_\alpha(\xi) = \frac{m(\xi)(1-e^{i(\alpha-\xi)})}{1-e^{i(\alpha-2\xi)}}$. Then the equation $[m]$ has a $\mathcal{C}_0$-solution if and only if $[m_\alpha]$ does. Furthermore, $\nu_0(m) = \nu_0(m_\alpha)$.*

Proof. Let T and T_α be the subdivision operators associated to the masks m and m_α, respectively.

Consider the operator $(P\psi)(x) = \sum_{k=0}^{N-2} p_k\psi(2x-k)$, where $p_0, \dots, p_{N-2}$ are the coefficients of the polynomial

$$p(\xi) = \sum_{k=0}^{N-2} p_k e^{-ik\xi} = \frac{m(\xi)}{1-e^{i(\alpha-2\xi)}}.$$

That is to say that in the frequency domain $\widehat{P\psi}(\xi) = \widehat{\psi}(\xi/2)p(\xi/2)$. It is clear that P is a continuous operator on $\mathcal{C}_0$. Furthermore, it preserves the subspace $\mathcal{L}$. Indeed, for any $\psi \in \mathcal{L}$ and $n \in \mathbb{Z}$ we have $\widehat{P\psi}(2\pi n) = \widehat{\psi}(\pi n)p(\pi n) = 0$. (If n is even, then $\widehat{\psi}(\pi n) = 0$; if n is odd, then $p(\pi n) = 0$, since the mask m satisfies (2.2).) Now observe that

$$PS_\alpha = T_\alpha, \quad S_\alpha P = T. \tag{4.2}$$

To prove this we apply (1.4) and get, consequently,

$$\widehat{PS_\alpha\psi}(\xi) = p(\xi/2)(1-e^{i(\alpha-\xi/2)})\widehat{\psi}(\xi/2) = m_\alpha(\xi/2)\widehat{\psi}(\xi/2) = \widehat{T_\alpha\psi}(\xi).$$

The equality $S_\alpha P = T$ can be proved in the same way.

Let $\psi \in \mathcal{C}_0$ be a solution of the equation $[m_\alpha]$. Since $T(S_\alpha\psi) = S_\alpha PS_\alpha\psi = S_\alpha T_\alpha\psi = S_\alpha\psi$, we see that the function $S_\alpha\psi$ is a solution of the equation $[m]$. Conversely, if a function $\varphi \in \mathcal{C}_0$ satisfies $T\varphi = \varphi$, then by (4.2) we have $\varphi \in \mathcal{C}_0\{\alpha\}$. Hence, by Lemma 4.1, the function $\psi = S_\alpha^{-1}\varphi$ is well defined and belongs to $\mathcal{C}_0$. Now, arguing as above, we obtain $T_\alpha\psi = \psi$.

From (4.2) it follows that $T^k = S_\alpha T_\alpha^{k-1}P$ for every $k \geq 1$. Therefore, since P and S_α are continuous and preserve the subspace $\mathcal{L}$, we see that $\nu_0(m) \geq \nu_0(m_\alpha)$.

Conversely, from the equality $T_{\alpha}^{k} = PT^{k-1}S_{\alpha}$ it follows that $\nu_0(m_{\alpha}) \geq \nu_0(m)$. Proposition 1 is proved. □

So using Proposition 1 we can consequently eliminate all symmetric roots and pass from the refinement equation with mask m to one with mask $R[m]$. The next step is to eliminate all cycles of the polynomial $R[m]$. In order to realize it we use the *matrix technique*, which was successfully applied in the study of subdivision processes [MP],[CDM], [DL1],[W],[E]. For a given refinement equation $[m]$ consider the two linear operators B_0 and B_1 acting on $\mathbb{C}^N$ and defined by $N \times N$ matrices as follows:

$$(B_0)_{ks} = c_{2k-s-1}, \quad (B_1)_{ks} = c_{2k-s}, \tag{4.3}$$

where c_j is the coefficient of (1.1) if $j \in \{0, 1, \ldots, N\}$, and $c_j = 0$ otherwise. As usual, we denote 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, and by $V^{\perp}$ the orthogonal complement of a subspace V in Euclidean space. Let us recall the notion of the joint spectral radius of finite-dimensional linear operators:

$$\hat{\rho}(A_1, A_2) = \lim_{n\to\infty} \max_{(d_1,\ldots,d_n)\in\{0,1\}^n} .\|A_{d_1}\cdots A_{d_n}\|^{1/n}.$$

See [RoS], [BW], [CH], [LW], [P1] for more details about the joint spectral radius.

We need the following two lemmas. The first one is a direct corollary of results of the works [DL2] and [CH]. The proof of the second one can be found in [HC] or [P1].

Lemma 4.2 (see [DL2], [CH]). *Let Δ be a finite family of real values such that the space $\mathcal{L}_{\Delta}$ is invariant with respect to the subdivision operator T; then*

$$\nu_0(m, \mathcal{L}_{\Delta}) = -\log_2 \hat{\rho}(B_0|_V, B_1|_V),$$

where

$$V = span\ \{(f(x), \ldots, f(x+N-1))^T \in \mathbb{C}^N \mid f \in \mathcal{L}_{\Delta}[0, N],\ x \in [0,1]\}.$$

In particular,

$$\nu_0(m) = \hat{\rho}(B_0|_W, B_1|_W),\ \textit{where}\ W = \left\{(x_1, \cdots, x_N)^T \in \mathbb{C}^N \mid \sum x_j = 0\right\}.$$

Lemma 4.3 (see [HC], [P1]). *Let A_0 and A_1 be linear operators acting on a finite-dimensional Euclidean space E. Suppose E_0 is a nontrivial common invariant subspace of these operators; then*

$$\hat{\rho}(A_0, A_1) = \max\Big\{\hat{\rho}(A_0|_{E_0}, A_1|_{E_0}), \hat{\rho}(A_0^*|_{E_0^{\perp}}, A_1^*|_{E_0^{\perp}})\Big\}.$$

Now we are able to realize the second step of the proof of Theorems 3.1 and 3.2.

Proposition 2. *Suppose a mask $m(\xi)$ possesses a cycle $\mathbf{b} = \{\beta_1, \ldots, \beta_n\}$. Denote by $\tilde{m}(\xi)$ the polynomial $m(\xi)/\prod_{k=1}^{n}(1 + e^{i(\beta_k - \xi)})$. Then the equation $[m]$ has a $\mathcal{C}_0$-solution if and only if $[\tilde{m}]$ does. Furthermore, $\nu_0(m) = min\{\nu_0(\tilde{m}), -\log_2 \rho_m(\mathbf{b})\}$.*

Proof. Consider the polynomial $q(\xi) = \prod_{k=1}^{n}(1 - e^{i(\beta_k - \xi)})$ and the corresponding operator $Q = S_{\beta_1} \circ \cdots \circ S_{\beta_n}$, which has the form $\widehat{Q\psi}(\xi) = \widehat{\psi}(\xi)q(\xi)$ in the frequency domain. It follows from Lemma 4.1 that Q maps the space $\mathcal{C}_0$ one-to-one into $\mathcal{C}_0\{\mathbf{b}\}$ and Q^{-1} is well defined and continuous on $\mathcal{C}_0\{\mathbf{b}\}$. Let T and $\tilde{T}$ be the subdivision operators associated to the masks m and $\tilde{m}$, respectively. For an arbitrary function $f \in \mathcal{C}_0\{\mathbf{b}\}$ we have

$$\widehat{Tf}(\xi)/q(\xi) = m(\xi/2)\hat{f}(\xi/2)/q(\xi) = \tilde{m}(\xi/2)\hat{f}(\xi/2)/q(\xi/2) \in \mathcal{H}.$$

Consequently, Tf is in $\mathcal{C}_0\{\mathbf{b}\}$ whenever $f \in \mathcal{C}_0\{\mathbf{b}\}$. This yields that the operator equality

$$\tilde{T} = Q^{-1}TQ \tag{4.4}$$

holds on the space $\mathcal{C}_0$. If a function $\psi \in \mathcal{C}_0$ satisfies the equality $\tilde{T}\psi = \psi$, then $\varphi = Q\psi$ satisfies $T\varphi = \varphi$. Conversely, assume that a function $\varphi \in \mathcal{C}_0$ satisfies $T\varphi = \varphi$. First let us show that φ belongs to $\mathcal{C}_0\{\mathbf{b}\}$. Using (1.2) we get

$$\widehat{\varphi}(\xi) = \widehat{\varphi}(0)\prod_{r=1}^{\infty} m\left(\frac{\xi}{2^r}\right) = \widehat{\varphi}(0)\prod_{r=1}^{\infty} \frac{q(\xi/2^{r-1})}{q(\xi/2^r)}\tilde{m}\left(\frac{\xi}{2^r}\right) = \frac{q(\xi)\widehat{\varphi}(0)}{q(0)}\prod_{r=1}^{\infty}\tilde{m}\left(\frac{\xi}{2^r}\right).$$

Since the function $\prod_{r=1}^{\infty}\tilde{m}(\frac{\xi}{2^r})$ is entire, it follows that $\varphi \in \mathcal{C}_0\{\mathbf{b}\}$, whence the function $\psi = Q^{-1}\varphi$ is well defined and obviously satisfies $\tilde{T}\psi = \psi$.

Now in order to prove the equality $\nu_0(m) = \min\ \{\nu_0(\tilde{m}), -\log_2\rho_m(\mathbf{b})\}$ we are going to use Lemmas 4.2 and 4.3. Let B_0 and B_1 be the linear operators acting in $\mathbb{C}^N$ and defined from (4.3). For arbitrary $t \in \mathbb{T}$ let us denote the vector $u(t) = (1, e^{it}, e^{2it}, \dots, e^{i(N-1)t})^T \in \mathbb{C}^N$. Further, define the following subspaces:

$$U = span\ \{u(\beta_1), \dots, u(\beta_n)\}, \quad W = u(0)^{\perp} = \left\{(x_1, \dots, x_N) \in \mathbb{C}^N \mid \sum x_k = 0\right\},$$

and

$$\tilde{W} = \{u(0), u(\beta_1), \dots, u(\beta_n)\}^{\perp}.$$

Finally, denote $A_i = B_i|_W,\ \tilde{A}_i = B_i|_{\tilde{W}}, i = 0, 1.$

From (4.4) it follows that the equality $T^k = Q\tilde{T}^kQ^{-1}$ holds on the space $\mathcal{L}_{\mathbf{b}}$ for any $k \geq 1$. This yields that $\nu_0(\tilde{m}) = \nu_0(m, \mathcal{L}_{\mathbf{b}})$. If we combine this with Lemma 4.2, we get $\nu_0(\tilde{m}) = -\log_2\hat{\rho}(\tilde{A}_0, \tilde{A}_1)$. Now it remains to prove the equality

$$\hat{\rho}(A_0, A_1) = \max\{\hat{\rho}(\tilde{A}_0, \tilde{A}_1), \rho_m(\mathbf{b})\}. \tag{4.5}$$

To do this observe the following property of operators B_0 and B_1:

(4.6)

$$B_0^* u(t) = \overline{m\left(\frac{t}{2}\right)}u\left(\frac{t}{2}\right) + \overline{m\left(\frac{t}{2}+\pi\right)}u\left(\frac{t}{2}+\pi\right), \quad t \in \mathbb{T},$$

$$B_1^* u(t) = e^{-\frac{it}{2}}\overline{m\left(\frac{t}{2}\right)}u\left(\frac{t}{2}\right) + e^{-i(\frac{t}{2}+\pi)}\overline{m\left(\frac{t}{2}+\pi\right)}u\left(\frac{t}{2}+\pi\right), \quad t \in \mathbb{T}.$$

(This can be easily shown by a direct calculation; see also [P1] or [CD2].) Therefore, for arbitrary $\beta_k \in \mathbf{b}$ the following hold:

$$B_0^* u(\beta_k) = \overline{m(\beta_{k-1})}u(\beta_{k-1}), \quad B_1^* u(\beta_k) = e^{-i\beta_{k-1}}\overline{m(\beta_{k-1})}u(\beta_{k-1}).$$

Therefore, for any $\beta_k \in \mathbf{b}$ and any set of indices $\{d_1, \dots, d_n\} \in \{0, 1\}^n$ we have

$$B_{d_1}^* \cdots B_{d_n}^* u(\beta_k) = e^{i\mu}\left(\prod_{j=1}^{n}\overline{m(\beta_j)}\right)u(\beta_k),$$

where $\mu \in \mathbb{T}$ depends on β and $d_1, \dots, d_n$. Since the vectors $\{u(\beta_k)\}_{k=1}^n$ form a basis of the space U, it follows that the operator $B^*_{d_1} \cdots B^*_{d_n}|_U$ is expressed in that basis by a diagonal matrix and moreover, the modulus of each diagonal entry of that matrix is equal to $|\prod_{j=1}^n m(\beta_j)| = (\rho_m(\mathbf{b}))^n$. This implies immediately that

$$\hat{\rho}(B^*_0|_U, B^*_1|_U) = \rho_m(\mathbf{b}). \tag{4.7}$$

If we apply Lemma 4.3 to the space W, its subspace $\tilde{W}$, and operators A_0, A_1 defined above, we obtain

$$\hat{\rho}(A_0, A_1) = \max\Big\{\hat{\rho}(\tilde{A}_0, \tilde{A}_1), \hat{\rho}(A^*_0|_H, A^*_1|_H)\Big\},$$

where H is the orthogonal complement of the subspace $\tilde{W}$ in the space W. Let us finally note that $A^*_i|_H = P_H B^*_i|_U P_H^{-1}, i = 0, 1$, where P_H is the operator of orthogonal projection from U to H. (Since the vectors $u(0), u(\beta_1), \dots, u(\beta_n)$ are linearly independent, it follows that P_H^{-1} is well defined on the space H.) Combining this with (4.7), we get

$$\hat{\rho}(A^*_0|_H, A^*_1|_H) = \hat{\rho}(B^*_0|_U, B^*_1|_U) = \rho_m(\mathbf{b}),$$

which completes the proof of Proposition 2. □

Suppose we have a subdivision process $\{m_0\}$; then we pass to the process $\{R[m_0]\}$ and, using Proposition 2, consequently eliminate all cycles of the mask $R[m_0]$. As a result we obtain the mask m_1 that has neither symmetric roots nor cycles. So we prove the following statement, which is a weaker version of Theorem 3.2.

Proposition 3. *For a given mask m_0 satisfying* (2.2) *let us denote*

$$m_1(\xi) = R[m_0](\xi) / \prod_{k=1}^{q} \prod_{\beta \in \mathbf{b}_k} (1 + e^{i(\beta - \xi)}),$$

where $\{\mathbf{b}_1, \dots, \mathbf{b}_q\}$ is the set of cycles of the polynomial $R[m_0]$ (counting with multiplicity). Then we have the following: the equation $[m_0]$ has a $\mathcal{C}_0$-solution if and only if $[m_1]$ does; furthermore,

$$\nu_0(m_0) = \min\{\nu_0(m_1), -\log_2 \rho_{m_0}(\mathbf{b}_1), \dots, -\log_2 \rho_{m_0}(\mathbf{b}_q)\}.$$

Thus Theorem 3.2 is proved for the case $l = 0$. Combining this with Theorem 2.1 we obtain Theorem 3.1 for the case $l = 0$.

Now it remains to realize the third step of the proof, i.e., to extend the statements of Theorems 3.1 and 3.2 from the case $l = 0$ to the general integer factor $l \geq 0$. To do this we introduce Proposition 4, which gives a method of factorization of refinement equations. Proposition 4 reduces the study of refinable functions and subdivision processes from the space $\mathcal{C}^l$ to $\mathcal{C}$.

Let us first remember the definition of the *cardinal B-spline*:

$$B_0(x) = \chi_{[0,1]}(x); \quad B_k(x) = [\chi_{[0,1]} * \cdots * \chi_{[0,1]}](x) \quad (k \text{ convolutions}).$$

For any $k \geq 0$ the cardinal B-spline B_k is a solution of the refinement equation with mask $(\frac{1+e^{-i\xi}}{2})^{k+1}$ (see, for instance, [Sc] or [DL2]).

Proposition 4. *Suppose m and m_0 are masks of refinement equations such that $m(\xi) = (\frac{1+e^{-i\xi}}{2})^l m_0(\xi)$, $l \geq 1$; then*

(a) *the equation* $[m]$ *has a* $\mathcal{C}_0^l$*-solution if and only if* $[m_0]$ *has a* $\mathcal{C}_0$*-solution. Moreover,* $\psi = S_0^{-l}\varphi^{(l)}$ *and* $\varphi = B_{l-1} * \psi$*, where* φ *and* ψ *are solutions of* $[m]$ *and* $[m_0]$*, respectively;* S_0 *is the difference operator:* $S_0 f(x) = f(x) - f(x-1)$.

(b) *The subdivision process* $\{m\}$ *converges in* $\mathcal{C}^l$ *if and only if* $\{m_0\}$ *converges in* $\mathcal{C}$*. Moreover,* $\nu(m) = \nu(m_0) + l$.

Proof. It follows from Lemma 4.1 that the mapping $S_0^l : \mathcal{C}_0 \to \mathcal{M}^{l-1}$ is a homeomorphism. Furthermore, for any $k \geq 0$ the mapping $S_0^l : \mathcal{M}^k \to \mathcal{M}^{k+l}$ is a homeomorphism. Now observe that for any $f \in \mathcal{C}_0$ and $g \in \mathcal{M}^{l-1}$ we have

$$T_0 f = 2^l S_0^{-l} T S_0^l f, \quad f \in \mathcal{C}_0,$$

$$Tg = 2^{-l} S_0^l T_0 S_0^{-l} g, \quad g \in \mathcal{M}^{l-1}, \tag{4.8}$$

where T and T_0 are the subdivision operators associated to the masks m and m_0, respectively. This immediately implies item (a). Further, from (4.8) it follows that $\nu_0(m, \mathcal{M}^{k+l}) = \nu_0(m_0, \mathcal{M}^k) + l$ for any admissible $k \geq 0$, i.e. whenever $k \leq \mathbf{l}(m_0)$. Therefore, $\nu_{k+l}(m) = \nu_k(m_0) + l$. Combining this with (3.1) we obtain item (b), which completes the proof of Proposition 4. □

Now to extend Theorems 3.1 and 3.2 from the case $l = 0$ it is sufficient to pass from the mask m to m_0 (applying Proposition 4) and note that $\rho_m(\mathbf{b}) = 2^{-l}\rho_{m_0}(\mathbf{b})$ for any cycle $\mathbf{b}$. This concludes the proof of the main theorems. □

Remark 5. The statement of item (a) of Proposition 4 generalizes the result [E, Theorem 2.2], which was obtained for refinement equations satisfying Cohen's criterion (see Remark 1).

Remark 6. It follows from results of the work [P2] that the statement of item (a) of Proposition 4 can be extended to general refinement equations, i.e., equations without condition (2.1). Namely, the following hold.

If an equation $[m]$ *has a* $\mathcal{C}_0^l$*-solution* $\varphi(x)$, $(l \geq 1)$*, then there exist dyadic rational values* $\gamma_1, \ldots, \gamma_r$ *(perhaps coinciding) such that* $\varphi = B_{l-1} * (S_{\gamma_1} \circ \cdots \circ S_{\gamma_r}\psi)$ *(and correspondingly* $\psi = S_0^{-l} \circ S_{\gamma_1}^{-1} \circ \cdots \circ S_{\gamma_r}^{-1}\varphi^{(l)}$*), where* ψ *is the* $\mathcal{C}_0$*-solution of the equation having the mask*

$$m_0(\xi) = \frac{m(\xi)}{[(1+e^{-i\xi})/2]^l} \prod_{k=1}^{r} \frac{1 - e^{i(2\pi\gamma_k - \xi)}}{1 - e^{i(2\pi\gamma_k - 2\xi)}}.$$

So the study of smooth refinable functions can be reduced to the study of continuous refinable functions (see [P2] for more details; see also [R1] and [C] for similar factorization theorems).

5. Generalized cycles. Theorems 3.1 and 3.2 are formulated in terms of cycles of the polynomial $R[m]$. It is easy to see that in general the sets of cycles of the polynomials m and $R[m]$ are different. The question arises, How can cycles of $R[m]$ be characterized by roots of m? In other words, we are going to reformulate the criterion of stability of subdivision operator in terms of zeros of its mask.

Let $p(\xi)$ be a given trigonometric polynomial. (Let us remember that we consider polynomials without positive powers.) Assume that p possesses a pair of symmetric roots $\{\alpha/2, \pi + \alpha/2\}$. The transfer from $p(\xi)$ to the polynomial $p_\alpha(\xi) = \frac{p(\xi)(1-e^{i(\alpha-\xi)})}{1-e^{i(\alpha-2\xi)}}$ is said to be a *transfer to the previous level.* The inverse transfer from p_α to p is a *transfer to the next level.* So the polynomial $R[p]$ is obtained from p by a sequence of transfers to the previous level.

To a given value $\alpha \in \mathbb{T}$ we assign a binary tree denoted in what follows by $\mathcal{T}_\alpha$. To every vertex of this tree we associate a value from $\mathbb{T}$ as follows: put α at the root, then put $\alpha/2$ and $\pi + \alpha/2$ at the vertices of the first level. (The *level* of the vertex is the distance from this vertex to the root. The root has level 0.) If a value γ is associated to a vertex on the nth level, then the values $\gamma/2$ and $\pi + \gamma/2$ are associated to its neighbors on the $(n+1)$st level. Thus there are the values $\frac{\alpha}{2^n} + \frac{2k\pi}{2^n}$, $k = 0, \dots, 2^n - 1$ on the nth level of the tree $\mathcal{T}_\alpha$. A set of vertices $\mathcal{A}$ of the tree $\mathcal{T}_\alpha$ is called a *minimal cut set* if every infinite path (all the paths are without backtracking) starting at the root includes exactly one element of $\mathcal{A}$. For instance, the one-element set $\mathcal{A} = \{root\}$ is a minimal cut set.

Definition 5.1. *A set $\{\beta_1, \dots, \beta_n\} \subset \mathbb{T}$ is called a generalized cycle of the polynomial $p(\xi)$ if the following hold:*

(a) *this set is cyclic, i.e., $\beta_{j+1} = 2\beta_j$ for all $j = 1, \dots, n$ (we set $\beta_{n+1} = \beta_1$);*

(b) *for any $j = 1, \dots, n$, the tree $\mathcal{T}_{\beta_j + \pi}$ possesses a minimal cut set that consists of roots of the polynomial p.*

Any (regular) cycle of $p(\xi)$ is also a generalized cycle. Indeed, in this case each minimal cut set A_j is the root of the corresponding tree $\mathcal{T}_{\beta_j+\pi}$. Now we establish a correlation between generalized cycles of the polynomial $p(\xi)$ and (regular) cycles of $R[p]$.

Proposition 5. (a) *Every cycle of the polynomial $R[p]$ is a generalized cycle of p.*

(b) *Every generalized cycle $\mathbf{b}$ of the polynomial p such that $\rho_p(\mathbf{b}) \neq 0$ is a cycle of $R[p]$.*

Proof. (a) Let $\mathbf{b} = \{\beta_1, \dots, \beta_n\}$ be a cycle of the polynomial $R[p]$. The polynomial p is obtained from $R[p]$ by a sequence of transfers to the next level. That sequence takes the root of the tree $\mathcal{T}_{\beta_j+\pi}$ to some minimal cut set $\mathcal{A}_j$ of this tree. Since $\beta_j + \pi$ is a root of $R[p]$, it follows that all elements of $\mathcal{A}_j$ are roots of p. So the set $\mathbf{b}$ is a generalized cycle for $p(\xi)$.

(b) Let $\mathbf{b} = \{\beta_1, \dots, \beta_n\}$ be a generalized cycle of the polynomial $p(\xi)$. Applying a suitable sequence of transfers to the previous level, we pass from the minimal cut sets $\mathcal{A}_1, \dots, \mathcal{A}_n$ to the roots $\beta_1 + \pi, \dots, \beta_n + \pi$ of the corresponding trees. Then we continue applying transfers to the previous level until we obtain the polynomial $R[p]$. If at some step we involve an element $\beta_j + \pi$ in this process, then the polynomial $p_1(\xi)$, which is obtained from the polynomial $p(\xi)$ by this step, has the pair of symmetric roots $\{\beta_j, \beta_j + \pi\}$. This implies that $\rho_{p_1}(\mathbf{b}) = 0$, and hence $\rho_p(\mathbf{b}) = 0$. Consider the opposite case. If the elements $\beta_1 + \pi, \dots, \beta_n + \pi$ are not involved, then each of them is a root of $R[p]$. Therefore, $\mathbf{b}$ is a cycle of $R[p]$. This completes the proof. □

Corollary 4. *If a polynomial $p(\xi)$ has no symmetric roots, then the set of its generalized cycles coincides with the set of its (regular) cycles.*

Corollary 5. *The set of all generalized cycles of a polynomial $p(\xi)$ is a union of the following two sets: the first one is the set of all cycles of $R[p]$; the second one consists of generalized cycles $\mathbf{b}$ such that $\rho_p(\mathbf{b}) = 0$.*

It follows from Propositions 2 and 4 that any cycle $\mathbf{b}$ such that $\rho_m(\mathbf{b}) = 0$ does not have influence on the convergence of the subdivision process $\{m\}$, i.e., $\nu(m) = \nu(\tilde{m})$ in terms of Proposition 2. Hence the criterion of convergence for subdivision processes can be formulated in terms of generalized cycles of mask. As a corollary we obtain the following main result of this section.

Corollary 6. *The statement of Theorem 3.1 remains true if the notion "a cycle of the polynomial $R[m]$" is replaced by "a generalized cycle of the mask m."*

Acknowledgment. The author is very grateful to Professor Ingrid Daubechies at Princeton University for useful discussions.

REFERENCES

[BW] M. A. Berger and Y. Wang, *Bounded semi-groups of matrices*, Linear Algebra Appl., 166 (1992), pp. 21–27.

[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, Boston 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. Collela and C. Heil, *Characterization of scaling functions:* I. *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. 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. 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.

[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, Birkhäuser, Basel, 1990, pp. 91–106.

[E] B. T. En, *Smoothness of wavelet and joint spectral radius*, J. Math. Sci. Univ. Tokyo, 5 (1998), pp. 241–256.

[H] C. Heil, *Some stability properties of wavelets and scaling functions*, in Wavelets and Their Applications, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 442, Kluwer, 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, 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.

[JW] R. Q. Jia and J. Wang, *Stability and linear independence associated with wavelet decomposition*, Proc. Amer. Math. Soc., 117 (1993), pp. 1115–1124.

[JS] Q. Jiang and Z. Shen, *On existence and weak-stability of matrix refinable functions*, Constr. Approx., 15 (1999), pp. 337–353.

[LW] J. C. Lagarias and Y. Wang, *The finiteness conjecture for the generalized spectral radius of a set of matrices*, Linear Algebra Appl., 214 (1995), pp. 17–42.

[MP] C. A. Micchelli and H. Prautzsch, *Uniform refinement of curves*, Linear Algebra Appl., 114/115 (1989), pp. 841–870.

[P1] V. Protasov, *The joint spectral radius and invariant sets of several linear operators*, Fundam. Prikl. Mat., 2 (1996), pp. 205–231.

[P2] V. Protasov, *A complete solution characterizing smooth refinable functions*, SIAM J. Math. Anal., 31, (2000) pp. 1332–1350.

[R1] A. Ron, *Factorization theorems for univariate splines on regular grids*, Israel J. Math., 70 (1990), pp. 48–68.

[R2] A. Ron, *Smooth refinable functions provide good approximation orders*, SIAM J. Math. Anal., 28 (1997), pp. 731–748.

[R3] A. Ron, *Wavelets and their associated operators*, in Approximation Theory IX, Vol. 2 (Nashville, TN, 1998), Innov. Appl. Math., Vanderbilt University Press, Nashville, 1998, pp. 283–317.

[RS] A. Ron and Z. Shen, *The Sobolev regularity of refinable functions*, J. Approx. Theory, 106 (2000), pp. 185–225.

[RoS] G. C. Rota and G. Strang, *A note on the joint spectral radius*, Nederl. Acad. Wetensch. Proc. Ser. A., 63 (1960), pp. 379–381.

[Sc] L. L. Schumaker, *Spline Functions: Basic Theory*, John Wiley, New York, 1981.

[V] L. I. Villemoes, *Wavelet analysis of refinement equations*, SIAM J. Math. Anal., 25 (1994), pp. 1433–1460.

[W] Y. Wang, *Two-scale dilation equations and the cascade algorithm*, Random Comput. Dynam., 3 (1995), pp. 289–307.

[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. 33, No. 2, pp. 461–476

UNIQUENESS FOR ELASTIC WAVE SCATTERING BY ROUGH SURFACES*

T. ARENS†

Abstract. We consider a two-dimensional elastic wave scattering problem for an unbounded surface represented as the graph of a $C^{1,\alpha}$ function. The total displacement is assumed to vanish on the surface. We present a new radiation condition, the upwards propagating radiation condition, for such problems based on a similar condition recently introduced for acoustic scattering problems. The relation between this radiation condition and more commonly used conditions is discussed. Subsequently we prove uniqueness of solution to the scattering problem under this radiation condition for a general class of incident fields, including plane and cylindrical waves.

Key words. elastic waves, scattering theory, rough surfaces, radiation condition, uniqueness

AMS subject classifications. 35J55, 73D25

PII. S0036141099359470

1. Introduction.

To date there appears to be little rigorous mathematical study of scattering problems for time harmonic elastic waves involving infinite rough surfaces. This paper is a first contribution to close this gap by proposing a precisely formulated radiation condition for a class of two-dimensional such problems and proving uniqueness for a problem in this class under this radiation condition.

The propagation of time harmonic waves with circular frequency ω in an elastic solid with Lamé constants μ, λ ($\mu > 0$, $\lambda + \mu \geq 0$), is governed by Hooke's law

$$\tau_{jk} = \lambda \operatorname{div} \mathbf{u} \, \delta_{jk} + \mu \left(\frac{\partial u_j}{\partial x_k} + \frac{\partial u_k}{\partial x_j} \right), \qquad j,k = 1,2,3, \tag{1}$$

and by the equations of motion

$$\sum_{k=1}^{3} \frac{\partial \tau_{jk}}{\partial x_k} + \omega^2 u_j = 0, \qquad j = 1,2,3. \tag{2}$$

Here, the vector field $\mathbf{u}$ denotes the displacements and τ denotes the stress tensor. Inserting the components of τ as given by (1) into the equations of motion (2) yields the Navier equation

$$\mu \, \Delta \mathbf{u} + (\lambda + \mu) \text{ grad div} \, \mathbf{u} + \omega^2 \, \mathbf{u} = 0. \tag{3}$$

All waves are assumed to be traveling in a half-space bounded by a surface invariant in the x_3-direction, on which all displacements are assumed to vanish. Because of this special geometry, the system of equations (3) separates into two parts, one describing compressional and vertically polarized shear waves, and the other describing horizontally polarised shear waves. Here we will consider only the first part; the scattering problem is treated as a problem of plane strain. Thus the problem is two-dimensional

*Received by the editors September 7, 1999; accepted for publication January 28, 2001; published electronically July 31, 2001. This work was supported by the European Commission through a Marie Curie Research Training Grant and by the UK Engineering and Physical Sciences Research Council.
http://www.siam.org/journals/sima/33-2/35947.html

†Department of Mathematical Sciences, Brunel University, Uxbridge UB8 3PH, UK (Tilo.Arens@brunel.ac.uk).

and, assuming that the boundary is the graph of a function f, the domain under consideration is $\Omega := \{\mathbf{x} \in \mathbb{R}^2 : x_2 > f(x_1)\}$. We will assume throughout this paper that f is bounded and that $\partial\Omega$ is Lyapunov.

This problem has many engineering applications, notably in seismology. In [10, 14, 15] the problem is considered in the case of a periodic traction-free surface and various numerical and analytical methods for computing the solution are presented. However, there appears to be no mathematically rigorous attempt to prove uniqueness and existence of solution until the author's recent work for the periodic case [1, 2].

In the case of acoustic waves, a lot of progress has been made lately in proving uniqueness and existence for both incident plane waves and incident cylindrical waves [4, 5, 6, 7, 8, 16]. The first step was to introduce a new radiation condition that will ensure uniqueness of solution for a wide range of incident fields, notably plane waves and cylindrical waves. In the present paper, these results will be generalized to the elastic wave case.

In section 2, we begin by introducing some notations and stating some results from linear elasticity theory used later in the paper. In section 3, we propose a new radiation condition for elastic wave scattering, suggested by the *upwards propagating radiation condition* (UPRC) for the Helmholtz equation [4], and we analyze its relation to other radiation conditions. In particular, we show that the new radiation condition is satisfied by solutions to the Navier equation satisfying Kupradze's radiation condition [13]. Finally, in section 4, a mathematical formulation of the scattering problem as a boundary value problem is given and uniqueness of solution to this problem is proved. Throughout the paper, reference will be made to some results on the regularity of solutions to the Navier equation that have been collected in the appendix.

2. Preliminaries. We will start by introducing notations and making some definitions that will be helpful subsequently. For any set $\mathcal{S} \subset \mathbb{R}^m$ ($m \in \mathbb{N}$) denote by $BC(\mathcal{S})$ the set of bounded and continuous, complex valued functions on $\mathcal{S}$, a Banach space under the supremum norm $\|\cdot\|_{\infty;\mathcal{S}}$. As an extension of the usual Hölder spaces $C^{k,\alpha}(\bar{D})$ ($k \in \mathbb{N} \cup \{0\}$) for bounded domains D with norm $\|\cdot\|_{k,\alpha;D}$, for unbounded domains $\mathcal{S} \subset \mathbb{R}^m$, we will introduce the sets

$$\mathcal{V}^{k,\alpha}(\mathcal{S}) := \{u : u \in C^{k,\alpha}(\bar{D}) \text{ for any domain } D \subset\subset \mathcal{S}\}$$

and

$$C^{k,\alpha}(\mathcal{S}) := \left\{ u \in \mathcal{V}^{k,\alpha}(\mathcal{S}) : \sup_{D \subset\subset \mathcal{S}} \|u\|_{k,\alpha;D} < \infty \right\}.$$

We also introduce a norm on $C^{k,\alpha}(\mathcal{S})$ by defining, for $u \in C^{k,\alpha}(\mathcal{S})$,

$$\|u\|_{k,\alpha;\mathcal{S}} := \sup_{D \subset\subset \mathcal{S}} \|u\|_{k,\alpha;D},$$

and we remark that $C^{k,\alpha}(\mathcal{S})$ is a Banach space with this norm.

The domain Ω under consideration is

$$\Omega := \{\mathbf{x} \in \mathbb{R}^2 : x_2 > f(x_1)\},$$

where we assume $f \in C^{1,\alpha}(\mathbb{R})$ to be real-valued.

We further let $S := \partial\Omega$ and, for any $A > 0$, $S(A) := \{\mathbf{x} \in S : |x_1| < A\}$. The normal $\mathbf{n}$ to S will always be assumed to be pointing out of Ω.

For $h \in \mathbb{R}$, define $U_h := \{\mathbf{x} \in \mathbb{R}^2 : x_2 > h\}$ and $T_h := \{\mathbf{x} \in \mathbb{R}^2 : x_2 = h\}$. Furthermore, let $D_h := \Omega \setminus \overline{U}_h$. We define $T_h(A)$ and $D_h(A)$ analogously to $S(A)$ and finally introduce $\gamma(h, A) := \{\mathbf{x} \in \Omega : |x_1| = A, x_2 < h\}$. Throughout, the normals on T_h and those on $T_h(A)$ and $\gamma(h, A)$ will be assumed to be pointing out of D_h and $D_h(A)$, as appropriate.

Throughout, all vectors and vector fields shall be denoted in bold print. For $\mathbf{y} = (y_1, y_2)^\top \in \mathbb{R}^2$ and $h \in \mathbb{R}$, define

$$\mathbf{y}'_h := (y_1, 2h - y_2)^\top \qquad \text{and} \qquad \mathbf{y}^\perp := (y_2, -y_1)^\top.$$

In addition to the usual differential operators grad$\cdot$ and div$\cdot$, we will make use of

$$\operatorname{grad}^\perp u := \left(\frac{\partial u}{\partial x_2}, -\frac{\partial u}{\partial x_1}\right)^\top \qquad \text{and} \qquad \operatorname{div}^\perp \mathbf{u} := \frac{\partial u_1}{\partial x_2} - \frac{\partial u_2}{\partial x_1}.$$

The differential operator in the Navier equation is abbreviated by Δ^* defined as

$$\Delta^* \mathbf{u} := \mu \, \Delta \, \mathbf{u} + (\lambda + \mu) \operatorname{grad} \operatorname{div} \mathbf{u}.$$

The fundamental solution in free-field conditions to the Helmholtz equation $\Delta \, u + k^2 u = 0$ will also play a role; it is given by

$$\Phi(\mathbf{x}, \mathbf{y}) := \frac{i}{4} H_0^{(1)}(k|\mathbf{x} - \mathbf{y}|), \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^2,$$

where $H_0^{(1)}$ denotes the Hankel function of order 0 and of the first kind.

The rest of this section will be devoted to stating some results from the linear theory of elastic wave propagation. First, let us note that any solution $\mathbf{u}$ to the Navier equation can be decomposed into uniquely identified compressional (or longitudinal) and shear (or transversal) components as

$$\mathbf{u} = \mathbf{u}_p + \mathbf{u}_s,$$

where

$$\mathbf{u}_p := -\frac{1}{k_p^2} \operatorname{grad} \operatorname{div} \mathbf{u} \quad \text{and} \quad \mathbf{u}_s := -\frac{1}{k_s^2} \operatorname{grad}^\perp \operatorname{div}^\perp \mathbf{u},$$

and the wave numbers k_p and k_s satisfy

$$k_p^2 = \frac{\omega^2}{2\mu + \lambda}, \qquad k_s^2 = \frac{\omega^2}{\mu}. \tag{4}$$

We find that $\mathbf{u}_p$ ($\mathbf{u}_s$) is a solution to the vector Helmholtz equation with $k = k_p$ ($k = k_s$).

Recalling Hooke's law (1), we follow Kupradze [13] in introducing a *generalised stress tensor* $\mathcal{P} = (\pi_{jk})$ by

$$\pi_{jk} := \tilde{\lambda} \operatorname{div} \mathbf{u} \, \delta_{jk} + \mu \frac{\partial u_j}{\partial x_k} + \tilde{\mu} \frac{\partial u_k}{\partial x_j},$$

where $\tilde{\lambda}$, $\tilde{\mu}$ are real numbers satisfying $\tilde{\lambda} + \tilde{\mu} = \lambda + \mu$. Given a curve $\Lambda \subset \mathbb{R}^2$ with a normal $\mathbf{n}$, the *generalised stress vector* on Λ is defined by

$$\mathbf{P}\mathbf{u} := \mathcal{P} \, \mathbf{n} = (\mu + \tilde{\mu}) \frac{\partial \mathbf{u}}{\partial \mathbf{n}} + \tilde{\lambda} \, \mathbf{n} \operatorname{div} \mathbf{u} - \tilde{\mu} \, \mathbf{n}^\perp \operatorname{div}^\perp \mathbf{u}.$$

Where it is important to distinguish between derivatives taken with respect to $\mathbf{x}$ and $\mathbf{y}$, the notations $\mathbf{P}^{(\mathbf{x})}\cdot$ and $\mathbf{P}^{(\mathbf{y})}\cdot$ will be used.

Similarly to Green's identities, there hold the generalized Betti formulae as follows.

Lemma 2.1. *Let $B \subseteq \mathbb{R}^2$ be a domain in which the divergence theorem holds. The normal on ∂B will be assumed to be pointing out of B. Then for vector fields $\mathbf{v} \in C^1(\bar{B})$ and $\mathbf{w} \in C^2(\bar{B})$ the first generalized Betti formula holds:*

$$\int_B \mathbf{v} \cdot \Delta^* \mathbf{w}\, d\mathbf{x} = \int_{\partial B} \mathbf{v} \cdot \mathbf{P}\mathbf{w}\, ds - \int_B \mathcal{E}_{a,b}(\mathbf{v}, \mathbf{w})\, d\mathbf{x}, \tag{5}$$

where the symmetric bilinear form $\mathcal{E}_{\tilde{\mu},\tilde{\lambda}}$ is given by

$$\begin{aligned}\mathcal{E}_{\tilde{\mu},\tilde{\lambda}}(\mathbf{v}, \mathbf{w}) := (2\mu + \lambda) \left(\frac{\partial v_1}{\partial x_1} \frac{\partial w_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} \frac{\partial w_2}{\partial x_2} \right) + \mu \left(\frac{\partial v_1}{\partial x_2} \frac{\partial w_1}{\partial x_2} + \frac{\partial v_2}{\partial x_1} \frac{\partial w_2}{\partial x_1} \right) \\ + \tilde{\lambda} \left(\frac{\partial v_1}{\partial x_1} \frac{\partial w_2}{\partial x_2} + \frac{\partial v_2}{\partial x_2} \frac{\partial w_1}{\partial x_1} \right) + \tilde{\mu} \left(\frac{\partial v_1}{\partial x_2} \frac{\partial w_2}{\partial x_1} + \frac{\partial v_2}{\partial x_1} \frac{\partial w_1}{\partial x_2} \right).\end{aligned}$$

For $\mathbf{v}$, $\mathbf{w} \in C^2(\bar{B})$, the third generalized Betti formula holds:

$$\int_B (\mathbf{v} \cdot \Delta^*\, \mathbf{w} - \mathbf{w} \cdot \Delta^*\, \mathbf{v})\, d\mathbf{x} = \int_{\partial B} (\mathbf{v} \cdot \mathbf{P}\mathbf{w} - \mathbf{w} \cdot \mathbf{P}\mathbf{v})\, ds. \tag{6}$$

Proof. The proof is as in Kupradze [13] for the three-dimensional case. □

In the definition and analysis of the new radiation condition, we will make heavy use of the elastic Green's tensors for free-field conditions and for a half-space with rigid boundary. The matrix of fundamental solutions for the Navier equation (3) which is the Green's tensor for free-field conditions is given by

$$\begin{aligned}\Gamma(\mathbf{x}, \mathbf{y}) := &\frac{i}{4\mu} H_0^{(1)}(k_s |\mathbf{x} - \mathbf{y}|) \\ &+ \frac{i}{4\omega^2} \nabla_x \nabla_x^\top \left(H_0^{(1)}(k_s |\mathbf{x} - \mathbf{y}|) - H_0^{(1)}(k_p |\mathbf{x} - \mathbf{y}|) \right).\end{aligned} \tag{7}$$

The Green's tensor for the half-space U_h with a rigid surface (i.e., the first boundary value problem) is given by

$$\Gamma_{D,h}(\mathbf{x}, \mathbf{y}) := \Gamma(\mathbf{x}, \mathbf{y}) - \Gamma(\mathbf{x}, \mathbf{y}'_h) + \mathbf{U}(\mathbf{x}, \mathbf{y}), \qquad \mathbf{x}, \mathbf{y} \in \overline{U_h},\ \mathbf{x} \neq \mathbf{y}, \tag{8}$$

where

$$\begin{aligned}&\mathbf{U}(\mathbf{x}, \mathbf{y}) \\ &= -\frac{i}{2\pi\omega^2} \Bigg\{ \int_{-\infty}^{\infty} \frac{\mathrm{e}^{i\gamma_p(x_2+y_2-2h)} - \mathrm{e}^{i(\gamma_p(x_2-h)+\gamma_s(y_2-h))}}{\gamma_p\gamma_s + t^2} \begin{pmatrix} -t^2\gamma_s & t^3 \\ t\gamma_p\gamma_s & -t^2\gamma_p \end{pmatrix} \mathrm{e}^{-iX_1 t}\, dt \\ &\qquad - \int_{-\infty}^{\infty} \frac{\mathrm{e}^{i\gamma_s(x_2+y_2-2h)} - \mathrm{e}^{i(\gamma_s(x_2-h)+\gamma_p(y_2-h))}}{\gamma_p\gamma_s + t^2} \begin{pmatrix} t^2\gamma_s & t\gamma_p\gamma_s \\ t^3 & t^2\gamma_p \end{pmatrix} \mathrm{e}^{-iX_1 t}\, dt \Bigg\}\end{aligned}$$

and $X_1 := x_1 - y_1$,

$$\gamma_p := \begin{cases} \sqrt{k_p^2 - t^2}, & k_p^2 \geq t^2, \\ i\sqrt{t^2 - k_p^2}, & k_p^2 < t^2, \end{cases} \qquad \gamma_s := \begin{cases} \sqrt{k_s^2 - t^2}, & k_s^2 \geq t^2, \\ i\sqrt{t^2 - k_s^2}, & k_s^2 < t^2. \end{cases}$$

The following theorem, which was essentially proved in [1], establishes the properties of $\Gamma_{D,h}$.

THEOREM 2.2.

(i) *For* $\mathbf{y} \in U_h$ *fixed,* $\Gamma_{D,h}(\cdot,\mathbf{y}) - \Gamma(\cdot,\mathbf{y}) \in [C^\infty(U_h) \cap C^1(\overline{U_h})]^{2\times 2}$ *and its columns are solutions to the Navier equation* (3) *in* $U_h \setminus \{\mathbf{y}\}$*.*

(ii) *For* $\mathbf{y} \in U_h$, $\mathbf{x} \in \partial U_h$, *there holds* $\Gamma_{D,h}(\mathbf{x},\mathbf{y}) = 0$*.*

(iii) *Let* $\mathbf{x}$, $\mathbf{y} \in U_h$, $\mathbf{x} \neq \mathbf{y}$*. Then*

$$\Gamma_{D,h}(\mathbf{x},\mathbf{y}) = \Gamma_{D,h}(\mathbf{y},\mathbf{x})^\top.$$

(iv) *For any* $\varepsilon > 0$, *the estimate*

$$\max_{j,k=1,2} |\Gamma_{D,h,jk}(\mathbf{x},\mathbf{y})| \leq \frac{\mathcal{H}(x_2 - h, y_2 - h)}{|x_1 - y_1|^{3/2}}$$

holds for all $\mathbf{x},\mathbf{y} \in U_h$, $|x_1 - y_1| \geq \varepsilon$, *where* $\mathcal{H} \in C(\mathbb{R}^2)$*.*

(v) *Let* $\Gamma^{(p)}_{D,h} := -k_p^{-2}\, \mathrm{grad}_{\mathbf{x}}\, \mathrm{div}_{\mathbf{x}}\, \Gamma_{D,h}$ *denote the longitudinal and let* $\Gamma^{(s)}_{D,h} := \Gamma_{D,h} - \Gamma^{(p)}_{D,h}$ *denote the transversal part of* $\Gamma_{D,h}$*. Then, for* $\mathbf{x}$, $\mathbf{y} \in U_h$, *and* $r := |\mathbf{x}-\mathbf{y}|$,

$$\Gamma^{(p)}_{D,h}(\mathbf{x},\mathbf{y}) = O(r^{-1/2}), \quad \frac{\partial \Gamma^{(p)}_{D,h}}{\partial r}(\mathbf{x},\mathbf{y}) - ik_p \Gamma^{(p)}_{D,h}(\mathbf{x},\mathbf{y}) = o(r^{-1/2}),$$
$$\Gamma^{(s)}_{D,h}(\mathbf{x},\mathbf{y}) = O(r^{-1/2}), \quad \frac{\partial \Gamma^{(s)}_{D,h}}{\partial r}(\mathbf{x},\mathbf{y}) - ik_s \Gamma^{(s)}_{D,h}(\mathbf{x},\mathbf{y}) = o(r^{-1/2})$$

uniformly in $\mathbf{x}$ *and* $\mathbf{y}$ *as* $r \to \infty$*.*

Proof. The proof follows from Theorems 2.1, 2.4, and 2.5 and Lemma 3.3 in [1]. □

By applying $\mathbf{P}\cdot$ to $\Gamma_{D,h}$, the matrix functions $\Pi^{(1)}_{D,h}$ and $\Pi^{(2)}_{D,h}$, defined by

$$\begin{array}{rcl} \Pi^{(1)}_{D,h,jk}(\mathbf{x},\mathbf{y}) & := & \left(\mathbf{P}^{(\mathbf{x})}(\Gamma_{D,h,\cdot k}(\mathbf{x},\mathbf{y}))\right)_j, \\ \Pi^{(2)}_{D,h,jk}(\mathbf{x},\mathbf{y}) & := & \left(\mathbf{P}^{(\mathbf{y})}(\Gamma_{D,h,j\cdot}(\mathbf{x},\mathbf{y}))^\top\right)_k, \end{array} \qquad j,k = 1,2,$$

are obtained. For these matrices, similar results to those for $\Gamma_{D,h}$ hold as follows.

THEOREM 2.3.

(i) *For* $\mathbf{y} \in U_h$, *the columns of* $\Pi^{(2)}_{D,h}(\cdot,\mathbf{y})$ *are solutions to the Navier equation* (3) *in* $U_h \setminus \{\mathbf{y}\}$*.*

(ii) *For* $\mathbf{x} \in U_h$, *the rows of* $\Pi^{(1)}_{D,h}(\mathbf{x},\cdot)$ *are solutions to the Navier equation* (3) *in* $U_h \setminus \{\mathbf{x}\}$*.*

(iii) *Theorem* 2.2 (iv) *and* (v) *hold with* $\Gamma_{D,h}$ *replaced by* $\Pi^{(1)}_{D,h}$ *and* $\Pi^{(2)}_{D,h}$, *respectively.*

(iv) *For* $\mathbf{x}$, $\mathbf{y} \in U_h$, $\mathbf{x} \neq \mathbf{y}$, *there holds*

$$\Pi^{(2)}_{D,h}(\mathbf{x},\mathbf{y}) = \Pi^{(1)}_{D,h}(\mathbf{y},\mathbf{x})^\top.$$

(v) *Let* $B \subset U_h$ *be a bounded domain in which the divergence theorem holds. Then any solution* $\mathbf{u} \in [C^2(B) \cap C^1(\bar{B})]^2$ *to the Navier equation* (3) *can be represented as*

$$\mathbf{u}(\mathbf{x}) = \int_{\partial B} \left\{ \Gamma_{D,h}(\mathbf{x},\mathbf{y})\, \mathbf{P}\mathbf{u}(\mathbf{y}) - \Pi^{(2)}_{D,h}(\mathbf{x},\mathbf{y})\, \mathbf{u}(\mathbf{y}) \right\} ds(\mathbf{y})$$

for all $\mathbf{x} \in B$*.*

Proof. The proof follows from Theorems 3.1 and 3.4 in [1]. □

Remark 2.4. We also note that from a representation of Γ in terms of Hankel functions of order 0 and 1 [12, formula (2.6)] together with Theorem 2.2 and the fact that $\Gamma_{D,h} - \Gamma$ remains bounded for $\mathbf{x}$, $\mathbf{y} \in U_h$, we see that there exists some constant $C > 0$ such that

$$\max_{j,k=1,2} |\Gamma_{D,h,jk}(\mathbf{x},\mathbf{y})| \leq C \,|1 + \log|\mathbf{x}-\mathbf{y}|\,| \tag{9}$$

for $\mathbf{x}$, $\mathbf{y} \in U_h$. As a consequence, together with an application of Theorem 2.2 (iv), we have for $h < \inf f$ and $h' > \sup f$ that

$$\sup_{\mathbf{x}\in D_{h'}} \int_S \max_{j,k=1,2} |\Gamma_{D,h,jk}(\mathbf{x},\mathbf{y})|^2 \, ds(\mathbf{y}) < \infty.$$

Remark 2.5. Similarly to (9), we also prove that $\Pi^{(2)}_{D,h}(\mathbf{x},\mathbf{y})$ remains bounded for $|\mathbf{x}-\mathbf{y}| \geq \varepsilon > 0$. Thus, using Theorem 2.2 (iv) and Lemma A.1, we see for $H' > H > h$ and any derivative with respect to $\mathbf{x}$, $\mathcal{G}$, of $\Pi^{(2)}_{D,h}$ that

$$\sup_{\mathbf{x}\in U_H\setminus U_{H'}} \int_{T_h} \max_{j,k=1,2} |\mathcal{G}_{jk}(\mathbf{x},\mathbf{y})| \, ds(\mathbf{y}) < \infty.$$

3. Radiation conditions for rough surface scattering. We will start this section by reviewing the acoustic case. Consider the Helmholtz equation

$$\Delta\, u + k^2 u = 0 \tag{10}$$

in some domain $G \subset \mathbb{R}^2$ such that for some $H \in \mathbb{R}$, $U_H \subset G$.

The standard radiation condition employed in problems of scattering by bounded obstacles for the Helmholtz equation was introduced by Sommerfeld as follows.

Definition 3.1. *A solution $u \in C^2(U_H) \cap L^\infty(U_H)$ to the Helmholtz equation* (10) *in U_H will be said to be* radiating *if*

$$u(\mathbf{x}) = O(r^{-1/2}), \qquad \frac{\partial u}{\partial r} - iku = o(r^{-1/2}),$$

uniformly in $\mathbf{x}/r$ as $r := |\mathbf{x}| \to \infty$.

Recently, a new radiation condition, the *upward propagating radiation condition* (UPRC) was introduced and successfully employed in a wide range of problems of scattering by unbounded rough surfaces and inhomogeneous layers [4, 6, 7, 8, 16] as follows.

Definition 3.2. *A solution $u : G \to \mathbb{C}$ to the Helmholtz equation* (10) *in $G \subset \mathbb{R}^2$ is said to satisfy UPRC, if, for some $H \in \mathbb{R}$ and $\phi \in L^\infty(T_H)$, $U_H \subset G$ and*

$$u(\mathbf{x}) = 2\int_{T_H} \frac{\partial \Phi}{\partial y_2}(\mathbf{x},\mathbf{y})\,\phi(\mathbf{y})\,ds(\mathbf{y}), \qquad \mathbf{x} \in U_H.$$

Let us now turn back to the elastic case. The radiation condition for the Navier equation (3) corresponding to Sommerfeld's radiation condition in the Helmholtz equation case is Kupradze's radiation condition, as given in the following definition.

Definition 3.3. *A solution $\mathbf{u} \in \left[C^3(U_H) \cap L^\infty(U_H)\right]^2$ to the Navier equation* (3) *in U_h will be said to be* radiating *if*

$$\mathbf{u}_p = O(r^{-1/2}), \quad \frac{\partial \mathbf{u}_p}{\partial r} - ik_p\mathbf{u}_p = o(r^{-1/2}),$$
$$\mathbf{u}_s = O(r^{-1/2}), \quad \frac{\partial \mathbf{u}_s}{\partial r} - ik_s\mathbf{u}_s = o(r^{-1/2}),$$

uniformly in $\mathbf{x}/r$ *as* $r := |\mathbf{x}| \to \infty$*, i.e., if both compressional and shear components of* $\mathbf{u}$ *are radiating solutions to the Helmholtz equation in* U_H*.*

The idea of the UPRC can be extended to the elastic case through the following definition.

DEFINITION 3.4. *A solution* $\mathbf{u} : G \to \mathbb{C}^2$ *to the Navier equation* (3) *in* $G \subset \mathbb{R}^2$ *is said to satisfy the UPRC, if, for some* $H \in \mathbb{R}$ *and* $\phi \in [L^\infty(T_H)]^2$, $U_H \subset G$ *and*

$$\mathbf{u}(\mathbf{x}) = \int_{T_H} \Pi^{(2)}_{D,H}(\mathbf{x},\mathbf{y})\,\phi(\mathbf{y})\,ds(\mathbf{y}), \qquad \mathbf{x} \in U_H. \tag{11}$$

Remark 3.5. Note that from Theorem 2.3 (iii) it follows that for arbitrary $\phi \in [L^\infty(T_h)]^2$ the integral in (11) exists as an improper integral.

Remark 3.6. Apparently the definition of the UPRC depends on the choice of the parameters $\tilde{\lambda}$ and $\tilde{\mu}$ in the definition of the generalized stresses. However, Theorem 3.7 below shows that the definition and the density ϕ itself are in fact independent of these numbers.

The following theorem characterizes the UPRC further and also establishes that it is satisfied by any radiating solution (see also [6, Theorem 2.9]).

THEOREM 3.7. *Given* $a \in \mathbb{R}$ *and* $\mathbf{u} : U_a \to \mathbb{C}^2$*, the following statements are equivalent:*

(i) $\mathbf{u} \in \left[C^2(U_a)\right]^2$, $\mathbf{u} \in [L^\infty(U_a \setminus U_H)]^2$ *for all* $H > a$, $\Delta^* \mathbf{u} + \omega^2\mathbf{u} = 0$ *in* U_a*, and* $\mathbf{u}$ *satisfies the UPRC in* U_a*.*

(ii) $\mathbf{u} \in \left[C^2(U_a)\right]^2$, $\mathbf{u} \in [L^\infty(U_a \setminus U_H)]^2$ *for all* $H > a$, $\Delta^* \mathbf{u} + \omega^2\mathbf{u} = 0$ *in* U_a*, and for some* $H > a$ *and* ϕ_1, $\phi_2 \in L^\infty(T_H)$,

$$\mathbf{u}(\mathbf{x}) = 2\ \mathrm{grad} \int\limits_{T_H} \frac{\partial \Phi_p}{\partial y_2}(\mathbf{x},\mathbf{y})\,\phi_1(\mathbf{y})\,ds(\mathbf{y}) + 2\ \mathrm{grad}^\perp \int\limits_{T_H} \frac{\partial \Phi_s}{\partial y_2}(\mathbf{x},\mathbf{y})\,\phi_2(\mathbf{y})\,ds(\mathbf{y})$$

for all $\mathbf{x} \in U_H$*, where* Φ_p *and* Φ_s *denote the fundamental solutions for the Helmholtz equation with* k *replaced by* k_p *and* k_s*, respectively.*

(iii) $\mathbf{u} \in [L^\infty(U_a \setminus U_H)]^2$ *for all* $H > a$ *and there exists a sequence* $(\mathbf{u}_n)$ *of radiating solutions such that* $\mathbf{u}_n(\mathbf{x}) \to \mathbf{u}(\mathbf{x})$ *uniformly on compact subsets of* U_a *and*

$$\sup_{\mathbf{x} \in U_H \setminus U_{h'}, n \in N} |\mathbf{u}_n(\mathbf{x})| < \infty \tag{12}$$

for all $H, h' \in \mathbb{R}$ *satisfying* $h' > H > a$*.*

(iv) $\mathbf{u}$ *satisfies* (11) *for* $H = a$ *and some* $\phi \in [L^\infty(T_a)]^2$*.*

(v) $\mathbf{u} \in [L^\infty(U_a \setminus U_H)]^2$ *for some* $H > a$ *and* $\mathbf{u}$ *satisfies* (11) *for each* $H > a$ *with* $\phi = \mathbf{u}|_{T_H}$*.*

(vi) $\mathbf{u} \in \left[C^2(U_a)\right]^2$, $\mathbf{u} \in [L^\infty(U_a \setminus U_H)]^2$ *for all* $H > a$, $\Delta^* \mathbf{u} + \omega^2\mathbf{u} = 0$ *in* U_a*, and for every* $H > a$ *and radiating solution in* U_a, $\mathbf{w}$*, such that the restrictions of* $\mathbf{w}$ *and* $\mathbf{Pw}$ *to* T_H *are in* $\left[L^1(T_H)\right]^2$*, there holds*

$$\int_{T_H} (\mathbf{u} \cdot \mathbf{Pw} - \mathbf{w} \cdot \mathbf{Pu})\,ds = 0. \tag{13}$$

Proof. (i) ⇒ (ii): With H chosen so that (11) holds, we introduce the functions

$$\begin{array}{rcl} \Psi_{p,k}(\mathbf{x},\mathbf{y}) & := & -\frac{1}{k_p^2}\,\mathrm{div}_{\mathbf{x}}\,\Pi^{(2)}_{D,H,\cdot k}(\mathbf{x},\mathbf{y}), \\ \Psi_{s,k}(\mathbf{x},\mathbf{y}) & := & -\frac{1}{k_s^2}\,\mathrm{div}^\perp_{\mathbf{x}}\,\Pi^{(2)}_{D,H,\cdot k}(\mathbf{x},\mathbf{y}), \end{array} \qquad k = 1, 2,$$

and rewrite $\mathbf{u}(\mathbf{x})$ for $\mathbf{x} \in U_H$ as

$$\begin{aligned}\mathbf{u}(\mathbf{x}) &= \mathbf{u}_p(\mathbf{x}) + \mathbf{u}_s(\mathbf{x}) \\ &= \operatorname{grad} \int_{T_H} \sum_{k=1}^{2} \Psi_{p,k}(\mathbf{x},\mathbf{y})\, \phi_k(\mathbf{y})\, ds(\mathbf{y}) \\ &\quad + \operatorname{grad}^{\perp} \int_{T_H} \sum_{k=1}^{2} \Psi_{s,k}(\mathbf{x},\mathbf{y})\, \phi_k(\mathbf{y})\, ds(\mathbf{y}). \end{aligned} \tag{14}$$

Limiting our attention to the first integral for the moment, we define

$$\begin{aligned}\mathbf{v}_N(\mathbf{x}) &= \int_{T_H(N)} \sum_{k=1}^{2} \Psi_{p,k}(\mathbf{x},\mathbf{y})\, \phi_k(\mathbf{y})\, ds(\mathbf{y}), \\ \mathbf{v}(\mathbf{x}) &= \int_{T_H} \sum_{k=1}^{2} \Psi_{p,k}(\mathbf{x},\mathbf{y})\, \phi_k(\mathbf{y})\, ds(\mathbf{y}).\end{aligned}$$

For $H' > H$, the vector fields $\mathbf{v}_N$ are solutions to the Helmholtz equation $\Delta \mathbf{v}_N + k_p^2 \mathbf{v}_N = 0$ in $U_{H'}$. By Theorem 2.2 (v) and two applications of Lemma A.1, we see that they are furthermore radiating in $U_{H'}$. By Theorem 2.2 (iv) together with Lemma A.1, there also holds $\mathbf{v}_N(\mathbf{x}) \to \mathbf{v}(\mathbf{x})$ uniformly on compact subsets of $U_{H'}$. For $h' > H'$, by Remark 2.5, we finally see that

$$\sup_{\mathbf{x} \in U_{H'} \setminus U_{h'}, n \in \mathbb{N}} |\mathbf{v}_N(\mathbf{x})| < \infty.$$

So by Theorem 2.1 in [6], $\mathbf{v}$ satisfies the UPRC for the Helmholtz equation (see Definition 3.2), which is the assertion. The argument for the second integral in (14) is identical.

(ii) ⇒ (iii): Set $\Psi_1 := -1/k_p^2 \operatorname{div} \mathbf{u}$ and $\Psi_2 := -1/k_s^2 \operatorname{div}^{\perp} \mathbf{u}$. Then (ii) implies that for all $\mathbf{x} \in U_H$ there holds

$$\begin{aligned}\Psi_1(\mathbf{x}) &= 2 \int_{T_H} \frac{\partial \Phi_p}{\partial y_2}(\mathbf{x},\mathbf{y})\, \phi_1(\mathbf{y})\, ds(\mathbf{y}), \\ \Psi_2(\mathbf{x}) &= 2 \int_{T_H} \frac{\partial \Phi_s}{\partial y_2}(\mathbf{x},\mathbf{y})\, \phi_2(\mathbf{y})\, ds(\mathbf{y}).\end{aligned}$$

From the equivalence of (i) and (ii) in Theorem 2.9 in [6], it follows that there exist sequences $(\Psi_j^{(n)})$ $(j = 1, 2)$ of radiating solutions to the Helmholtz equation with $k = k_p$ and $k = k_s$, respectively, such that $\Psi_j^{(n)}(\mathbf{x}) \to \Psi_j(\mathbf{x})$ uniformly on compact subsets of U_a and

$$\sup_{\mathbf{x} \in U_a \setminus U_h, n \in N, j=1,2} |\Psi_j^{(n)}(\mathbf{x})| < \infty$$

for all $h > a$. Set

$$\mathbf{u}_n(\mathbf{x}) := \operatorname{grad}\, \Psi_1^{(n)}(\mathbf{x}) + \operatorname{grad}^{\perp}\, \Psi_2^{(n)}(\mathbf{x}).$$

Lemma A.1 then implies (12) and that $\mathbf{u}_n(\mathbf{x})$ converges to $\mathbf{u}(\mathbf{x})$ uniformly on compact subsets of U_a.

(iii) ⇒ (vi): Suppose $H > a$ and set $D := U_H \cap B_R(0)$ for some $R > H$, where $B_R(0)$ denotes the open ball with center 0 and radius R. Further assume $\mathbf{w}$ to be a radiating solution in U_a, such that the restrictions of $\mathbf{w}$ and $\mathbf{Pw}$ to T_H are in $\left[L^1(T_H)\right]^2$. Then

$$\int_{\partial D} \{\mathbf{u}_n \cdot \mathbf{Pw} - \mathbf{w} \cdot \mathbf{Pu}_n\}\, ds = 0$$

follows from the third generalized Betti formula (6). Letting $R \to \infty$ and using the fact that $\mathbf{w}$ and $\mathbf{u}_n$ are radiating solutions to the Navier equation, we conclude that

$$\int_{T_H} \{\mathbf{u}_n \cdot \mathbf{Pw} - \mathbf{w} \cdot \mathbf{Pu}_n\}\, ds = 0.$$

Taking the limit as $n \to \infty$, recalling (12) and using Theorem 2.2 (iv) and Lemma A.1, we see that (13) holds. The remaining assertion follows from Corollary A.2.

(vi) ⇒ (i),(v): It suffices to show that (11) holds for all $H > a$ with $\phi = \mathbf{u}|_{T_H}$.

Given $H > a$ and $\mathbf{x} \in U_H$, choose $h', A \in \mathbb{R}$, with $h' > x_2 > H$ and $A > |x_1|$. Set $B := \{\mathbf{y} \in U_H \setminus \overline{U}_{h'} : |y_1| < A\}$. Then, by Theorem 2.3 (v),

$$\mathbf{u}(\mathbf{x}) = \int_{\partial B} \left\{\Gamma_{D,H}(\mathbf{x},\mathbf{y})\, \mathbf{Pu}(\mathbf{y}) - \Pi^{(2)}_{D,H}(\mathbf{x},\mathbf{y})\, \mathbf{u}(\mathbf{y})\right\} ds(\mathbf{y}).$$

Letting $A \to \infty$ and recalling $\mathbf{u} \in [L^\infty(U_a \setminus U_{h'})]^2$ as well as Theorem 2.2 (ii) and (iv) and Theorem 2.3 (iii), we obtain that

$$\begin{aligned}\mathbf{u}(\mathbf{x}) = &\int_{T_{h'}} \left\{\Gamma_{D,H}(\mathbf{x},\mathbf{y})\, \mathbf{Pu}(\mathbf{y}) - \Pi^{(2)}_{D,H}(\mathbf{x},\mathbf{y})\, \mathbf{u}(\mathbf{y})\right\} ds(\mathbf{y})\\ &+ \int_{T_H} \Pi^{(2)}_{D,H}(\mathbf{x},\mathbf{y})\, \mathbf{u}(\mathbf{y})\, ds(\mathbf{y}).\end{aligned}$$

By applying (13) with $\mathbf{w}$ equal to each of the rows of $\Gamma_{D,H}(\mathbf{x},\cdot)$ in turn, the integral over $T_{h'}$ is seen to vanish.

(v) ⇒ (iv): Introducing, for $\alpha \in \mathbb{R}$, the mapping

$$\eta_\alpha(\mathbf{z}) := (z_1, z_2 + \alpha)^\top,$$

we have from (v) that

$$\mathbf{u}(\mathbf{x}) = \int_{T_a} \Pi^{(2)}_{D,H}(\mathbf{x}, \eta_{H-a}(\mathbf{z}))\, \mathbf{u}(\eta_{H-a}(\mathbf{z}))\, ds(\mathbf{z}), \qquad \mathbf{x} \in U_H. \tag{15}$$

As $\mathbf{u} \in [L^\infty(U_a \setminus U_H) \cap C(U_a)]^2$ for some $H > a$, the densities $\mathbf{u}(\eta_{H-a}(\cdot))$ are all in some ball in $[L^\infty(T_a)]^2$ for H close enough to a. Recalling that the unit ball in $[L^\infty(T_a)]^2$ is weak$*$ sequentially compact, there thus exists a sequence (H_n) with $H_n \to a$ and $\mathbf{u}(\eta_{H_n-a}(\cdot)) \to \phi \in [L^\infty(T_a)]^2$. Taking the limit as $H \to a$, through this sequence in (15) we now conclude that (11) holds for $H = a$ with this ϕ.

(iv) ⇒ (iii): As (11) is satisfied with $h = a$, it follows from Theorem 2.3 (iii) that

$$|\mathbf{u}(\mathbf{x})| \le \|\phi\|_\infty\, g(x_2), \qquad \mathbf{x} \in U_a, \tag{16}$$

where $g \in C(\mathbb{R})$. Setting

$$\mathbf{u}_n(\mathbf{x}) := \int_{T_a(n)} \Pi^{(2)}_{D,a}(\mathbf{x},\mathbf{y})\, \phi(\mathbf{y})\, ds(\mathbf{y}), \qquad \mathbf{x} \in U_a,$$

$\mathbf{u} \in [L^\infty(U_a \setminus U_h)]^2$ for all $h > a$ and (12) follows from (16). That $\mathbf{u}_n(\mathbf{x})$ converges to $\mathbf{u}(\mathbf{x})$ uniformly on compact subsets of U_a, and that $\mathbf{u}_n$ is radiating, is also easily seen from Theorem 2.3 (iii). $\square$

Remark 3.8. In the case of a periodic boundary, one usually imposes a radiation condition using the Rayleigh expansion [2, 10, 14, 15]. Assume f to be 2π-periodic. Then $\mathbf{u} \in BC(\Omega)$ is said to satisfy the Rayleigh expansion radiation condition (RERC) if, for $x_2 > \max f$, it has an expansion of the form

$$\mathbf{u}(\mathbf{x}) = \sum_{n\in\mathbb{Z}} \left\{ u_{p,n} \begin{pmatrix} \alpha_n \\ \beta_n \end{pmatrix} \mathrm{e}^{i(\alpha_n x_1 + \beta_n x_2)} + u_{s,n} \begin{pmatrix} \gamma_n \\ -\alpha_n \end{pmatrix} \mathrm{e}^{i(\alpha_n x_1 + \gamma_n x_2)} \right\},$$

where $\alpha \in \mathbb{R}$, $\alpha \neq 0$, $u_{p,n}$, $u_{s,n} \in \mathbb{C}$ $(n \in \mathbb{Z})$, $\alpha_n := \alpha + n$,

$$\beta_n := \begin{cases} \sqrt{k_p^2 - \alpha_n^2}, & \alpha_n^2 \le k_p^2, \\ i\sqrt{\alpha_n^2 - k_p^2}, & \alpha_n^2 > k_p^2, \end{cases} \qquad \gamma_n := \begin{cases} \sqrt{k_s^2 - \alpha_n^2}, & \alpha_n^2 \le k_s^2, \\ i\sqrt{\alpha_n^2 - k_s^2}, & \alpha_n^2 > k_s^2. \end{cases}$$

A field $\mathbf{u}$ satisfying the RERC is quasi-periodic with phase-shift α in $U_{\max f}$; that is, for all $\mathbf{x} = (x_1, x_2)^\top \in U_{\max f}$,

$$\mathbf{u}(x_1 + 2\pi, x_2) = \mathrm{e}^{i\alpha 2\pi}\, \mathbf{u}(x_1, x_2).$$

From Remark 2.14 in [6] and the equivalence of statements (i) and (ii) in Theorem 3.7, it follows that any bounded solution to the Navier equation $\mathbf{u}$ in Ω that satisfies the UPRC in Ω and is quasi-periodic in Ω with phase-shift α also satisfies the RERC. Conversely, by the same arguments, a bounded, quasi-periodic solution to the Navier equation in Ω, satisfying the RERC, also satisfies the UPRC.

4. Uniqueness for the scattering problem with a rigid rough surface.

We will now address the goal of this paper, to prove uniqueness of solution to the scattering problem for a rigid rough surface. Let $\mathbf{u}^{inc}$ denote the incident field. We require only that $\mathbf{u}^{inc}$ is a solution to the Navier equation in some neighborhood of $S = \partial\Omega$ and that $\mathbf{g} := \mathbf{u}^{inc}|_S \in BC(S)$. The problem is then to find the scattered field $\mathbf{u}$ so that the total field $\mathbf{u}^{inc} + \mathbf{u}$ vanishes on S. Mathematically, this scattering problem will be formulated as the following boundary value problem.

Problem 4.1. Find a vector field $\mathbf{u} \in [C^2(\Omega) \cap C(\overline{\Omega}) \cap H^1_{loc}(\bar{\Omega})]^2$ that satisfies

1. the Navier equation $\Delta^* \mathbf{u} + \omega^2 \mathbf{u} = 0$ in Ω,
2. the Dirichlet boundary condition $\mathbf{u} = \mathbf{g}$ on S for some vector field $\mathbf{g} \in [BC(S)]^2$,
3. the vertical growth rate condition

$$\sup_{\mathbf{x}\in\Omega} x_2^\beta \, |\mathbf{u}(\mathbf{x})| < \infty \tag{17}$$

for some $\beta \in \mathbb{R}$, and

4. the UPRC in Ω.

Remark 4.2. A solution of Problem 4.1 satisfies statement (i) of Theorem 3.7 with any $a > \sup f$.

From Remark 3.8 we see that, in the case when f is periodic and the Dirichlet data $\mathbf{g}$ is quasi-periodic, Problem 4.1 reduces to the diffraction grating problem considered in [1, 2], if we assume additionally that the solution $\mathbf{u}$ is quasi-periodic. The diffraction grating problem was shown in [1, 2] to be uniquely solvable. Thus we know that

Problem 4.1 admits solutions, at least in the case when f is periodic and $\mathbf{g}$ is quasi-periodic. We now show that Problem 4.1 has at most one solution in every case. This result implies for the diffraction grating problem that the additional assumption on the solution of being quasi-periodic can in fact be dropped.

In all of what follows, let h denote a real number with $h < \inf f$. The first step in the uniqueness proof will be the following representation theorem.

THEOREM 4.3. *Let* $\mathbf{u}$ *be a solution to Problem* 4.1 *with* $\mathbf{g} \equiv 0$. *Then* $\mathbf{u} \in [C^1(\bar{\Omega})]^2$, *its first derivatives are bounded in* D_H *for any* $H > \sup f$, *and*

$$\mathbf{u}(\mathbf{x}) = \int_S \Gamma_{D,h}(\mathbf{x},\mathbf{y})\, \mathbf{P}\mathbf{u}(\mathbf{y})\, ds(\mathbf{y})$$

for all $\mathbf{x} \in \Omega$.

Proof. That $\mathbf{u} \in [C^1(\bar{\Omega})]^2$ and its first derivatives are bounded in D_H, $H > \sup f$, is a consequence of Theorem A.3. For $\mathbf{x} \in \Omega$, choose $h', A \in \mathbb{R}$, with $h' > \max\{x_2, \sup f\}$ and $A > |x_1|$. Then, by Theorem 2.3 (v), there holds

$$\mathbf{u}(\mathbf{x}) = \int_{\partial D_{h'}(A)} \left\{ \Gamma_{D,h}(\mathbf{x},\mathbf{y})\, \mathbf{P}\mathbf{u}(\mathbf{y}) - \Pi^{(2)}_{D,h}(\mathbf{x},\mathbf{y})\, \mathbf{u}(\mathbf{y}) \right\} ds(\mathbf{y}).$$

By applying Lemma A.1, we see that the growth rate condition (17) for $\mathbf{u}$ also holds for any first derivative of $\mathbf{u}$. Letting $A \to \infty$ and recalling Theorems 2.2 (iv) and 2.3 (iii) then yields

$$\mathbf{u}(\mathbf{x}) = \int_{T_{h'}} \left\{ \Gamma_{D,h}(\mathbf{x},\mathbf{y})\, \mathbf{P}\mathbf{u}(\mathbf{y}) - \Pi^{(2)}_{D,h}(\mathbf{x},\mathbf{y})\, \mathbf{u}(\mathbf{y}) \right\} ds(\mathbf{y}) \\ + \int_S \Gamma_{D,h}(\mathbf{x},\mathbf{y})\, \mathbf{P}\mathbf{u}(\mathbf{y})\, ds(\mathbf{y}).$$

The proof is now completed by recalling Remark 4.2 and the equivalence of (i) and (vi) in Theorem 3.7, by which the integral over $T_{h'}$ vanishes. □

Let us now introduce some functionals that will be of importance in the following arguments. Let $h' > \sup f$, $A > 0$, and $\mathbf{u} \in C^1(\bar{\Omega})$. We define

$$\begin{aligned} I(h',A)[\mathbf{u}] &:= \int_{T_{h'}(A)} \Bigg\{ (2\mu+\lambda) \left(\left| \frac{\partial u_2}{\partial x_2} \right|^2 - \left| \frac{\partial u_1}{\partial x_1} \right|^2 \right) \\ &\qquad + \mu \left(\left| \frac{\partial u_1}{\partial x_2} \right|^2 - \left| \frac{\partial u_2}{\partial x_1} \right|^2 \right) + \omega^2 |\mathbf{u}|^2 \Bigg\} ds, \\ J_1(h',A)[\mathbf{u}] &:= 2\,\mathrm{Re} \int_{\gamma(h',A)} \frac{\partial \bar{\mathbf{u}}}{\partial x_2} \cdot \mathbf{P}\mathbf{u}\, ds, \\ J_2(h',A)[\mathbf{u}] &:= \mathrm{Im} \int_{\gamma(h',A)} \bar{\mathbf{u}} \cdot \mathbf{P}\mathbf{u}\, ds, \\ K(h',A)[\mathbf{u}] &:= \mathrm{Im} \int_{T_{h'}(A)} \bar{\mathbf{u}} \cdot \mathbf{P}\mathbf{u}\, ds. \end{aligned}$$

Recall the assumptions on the direction of the normal vectors in section 2. The following lemma is of fundamental importance.

LEMMA 4.4. *Suppose* $\mathbf{u}$ *satisfies statement* (ii) *in Theorem* 3.7 *with* $H > \sup f$ *and some densities* $\phi_j \in L^2(T_H) \cap L^\infty(T_H)$ $(j = 1, 2)$. *Then, for all* $h' > H$, *there* *holds*

$$I(h',\infty)[\mathbf{u}] \le 2k_s\, K(h',\infty)[\mathbf{u}].$$

Proof. Choose $H > \sup f$ so that the representation for $\mathbf{u}$ in U_H according to statement (ii) of Theorem 3.7 holds. Then the argument presented for the derivation of (29) in [4] yields

$$\mathbf{u}(\mathbf{x}) = \frac{i}{2\pi}\int_{-\infty}^{\infty} \tilde{\phi}_1(t)\begin{pmatrix} t \\ \gamma_p\end{pmatrix} \mathrm{e}^{i(tx_1+\gamma_p x_2)} + \tilde{\phi}_2(t)\begin{pmatrix}\gamma_s \\ -t\end{pmatrix}\mathrm{e}^{i(tx_1+\gamma_s x_2)}\, dt, \qquad \mathbf{x}\in U_H,$$

where $\tilde{\phi}_1(t) := \mathrm{e}^{-i\gamma_p H}\hat{\phi}_1(t)$, $\tilde{\phi}_2(t) := \mathrm{e}^{-i\gamma_s H}\hat{\phi}_2(t)$, and $\hat{\phi}_j$ denotes the Fourier transform of $\phi_j(y_1,y_2)$ with respect to y_1 $(j=1,2)$.

By an application of Parseval's theorem we derive from this representation, for any $h' > H$, by long but straightforward calculations that

$$I(h',\infty)[\mathbf{u}] = 2\omega^2\left\{\int_{-k_p}^{k_p} |\tilde{\phi}_1|^2\gamma_p^2\, dt + \int_{-k_s}^{k_s}|\tilde{\phi}_2|^2\gamma_s^2\, dt\right\}.$$

On the other hand, a similar calculation shows that

$$K(h',\infty)[\mathbf{u}] = \omega^2\left\{\int_{-k_p}^{k_p} |\tilde{\phi}_1|^2\gamma_p\, dt + \int_{-k_s}^{k_s}|\tilde{\phi}_2|^2\gamma_s\, dt\right\}.$$

The assertion is now proven by noting $k_p \ge \gamma_p$ on $[-k_p,k_p]$, $k_s\ge\gamma_s$ on $[-k_s,k_s]$, and $k_s > k_p$. □

Another, simpler relation involving these functionals is stated in the following lemma.

LEMMA 4.5. *Let* $\mathbf{u}$ *be a solution to Problem* 4.1 *with* $\mathbf{g}\equiv 0$. *Further assume* $h' > \max f$ *and* $A > 0$. *Then*

$$K(h',A)[\mathbf{u}] = -J_2(h',A)[\mathbf{u}].$$

Proof. For the proof apply the third generalized Betti formula (6) to $\mathbf{u}$ and $\bar{\mathbf{u}}$ in $D_{h'}(A)$. □

Assume now that $\mathbf{u}$ is a solution to Problem 4.1 with $\mathbf{g}\equiv 0$. As $\mathbf{u}$ and its tangential derivatives vanish on S, $\mathbf{Pu}$ has the simple form

$$\mathbf{Pu} = \mu\frac{\partial\mathbf{u}}{\partial\mathbf{n}} + (\lambda+\mu)\,\mathbf{n}\,\mathrm{div}\,\mathbf{u} \qquad \text{on } S. \tag{18}$$

We thus conclude that

$$\int_{S(A)} \frac{\partial\bar{\mathbf{u}}}{\partial x_2}\cdot\mathbf{Pu}\, ds = \int_{S(A)}\left\{\mu n_2\left|\frac{\partial\mathbf{u}}{\partial\mathbf{n}}\right|^2 + (\lambda+\mu)n_2|\mathrm{div}\,\mathbf{u}|^2\right\} ds \tag{19}$$

for any $A>0$. For any $h' > \sup f$, by the first Betti formula (5) there also holds

$$\int_{\partial D_{h'}(A)} \frac{\partial\bar{\mathbf{u}}}{\partial x_2}\cdot\mathbf{Pu}\, ds = \int_{D_{h'}(A)} \mathcal{E}_{\tilde{\mu},\tilde{\lambda}}\left(\frac{\partial\bar{\mathbf{u}}}{\partial x_2},\mathbf{u}\right) - \omega^2\frac{\partial\bar{\mathbf{u}}}{\partial x_2}\cdot\mathbf{u}\, d\mathbf{x}.$$

By an integration by parts we thus conclude that

$$2\,\mathrm{Re}\int_{\partial D_{h'}(A)}\frac{\partial\bar{\mathbf{u}}}{\partial x_2}\cdot\mathbf{P}\mathbf{u}\,ds=\int_{T_{h'}(A)}\mathcal{E}_{\tilde{\mu},\tilde{\lambda}}(\bar{\mathbf{u}},\mathbf{u})-\omega^2|\mathbf{u}|^2\,ds$$
$$+\int_{S(A)}\left\{\mu n_2\left|\frac{\partial\mathbf{u}}{\partial\mathbf{n}}\right|^2+(\lambda+\mu)n_2|\mathrm{div}\,\mathbf{u}|^2\right\}ds. \tag{20}$$

Combining (19) and (20) now yields

$$\int_{S(A)}\left\{\mu n_2\left|\frac{\partial\mathbf{u}}{\partial\mathbf{n}}\right|^2+(\lambda+\mu)n_2|\mathrm{div}\,\mathbf{u}|^2\right\}ds$$
$$=\mathrm{Re}\int_{T_{h'}(A)}\left\{\mathcal{E}_{\tilde{\mu},\tilde{\lambda}}(\bar{\mathbf{u}},\mathbf{u})-2\frac{\partial\bar{\mathbf{u}}}{\partial x_2}\cdot\mathbf{P}\mathbf{u}-\omega^2|\mathbf{u}|^2\right\}ds-J_1(h',A)[\mathbf{u}].$$

It is also not difficult to see that

$$\mathrm{Re}\int_{T_{h'}}\left\{\mathcal{E}_{\tilde{\mu},\tilde{\lambda}}(\bar{\mathbf{u}},\mathbf{u})-2\frac{\partial\bar{\mathbf{u}}}{\partial x_2}\cdot\mathbf{P}\mathbf{u}-\omega^2|\mathbf{u}|^2\right\}ds=-I(h',A)[\mathbf{u}],$$

so we finally conclude that

$$0\le-\int_{S(A)}\left\{\mu n_2\left|\frac{\partial\mathbf{u}}{\partial\mathbf{n}}\right|^2+(\lambda+\mu)n_2|\mathrm{div}\,\mathbf{u}|^2\right\}ds$$
$$=I(h',A)[\mathbf{u}]+J_1(h',A)[\mathbf{u}]. \tag{21}$$

The rest of the derivation of the uniqueness result is now a rather straightforward adaptation of the method presented in [7] for the Helmholtz equation case. Let us introduce the vector fields $\mathbf{v}_A$ defined for $A>0$ by

$$\mathbf{v}_A(\mathbf{x}):=\int_{S(A)}\Gamma_{D,h}(\mathbf{x},\mathbf{y})\,\mathbf{P}\mathbf{u}(\mathbf{y})\,ds(\mathbf{y}).$$

Using the Cauchy–Schwartz inequality and Theorem 2.2 (iv), we find that $\mathbf{v}_A|_{h'}\in[L^2(T_{h'})\cap BC(T_{h'})]^2$ for all $h'>\sup f$. As $\mathbf{v}_A$ is a radiating solution to the Navier equation for every $A\in\mathbb{R}$, it is seen to satisfy statement (iii) of Theorem 3.7 and thus also statement (ii) of that theorem. Thus, by Lemma 4.4,

$$I(h',\infty)[\mathbf{v}_A]\le 2k_s\,K(h',\infty)[\mathbf{v}_A]. \tag{22}$$

Now set $w(x_1):=|\mathbf{P}\mathbf{u}(x_1,f(x_1))|$ for $x_1\in\mathbb{R}$. Then

$$\int_{-A}^{A}|w(x_1)|^2\,dx_1\le\int_{S(A)}|\mathbf{P}\mathbf{u}|^2\,ds\le(1+\|f'\|^2_{\infty;\mathbb{R}})^{1/2}\int_{-A}^{A}|w(x_1)|^2\,dx_1 \tag{23}$$

follows. Using Theorem 2.2 (iv) and Lemma A.1 we obtain the estimates

$$|\Gamma_{D,h}(\mathbf{x},\mathbf{y})|,\ \left|\frac{\partial}{\partial x_j}\Gamma_{D,h}(\mathbf{x},\mathbf{y})\right|\le C(1+|x_1-y_1|)^{-3/2},\qquad j=1,2,$$

for $\mathbf{x} \in T_{h'}$, $\mathbf{y} \in S$, where C is some positive constant depending only on h' and h. This yields the estimates

$$\begin{aligned} |\mathbf{v}_A(\mathbf{x})|,\ \left|\tfrac{\partial \mathbf{v}_A}{\partial x_j}(\mathbf{x})\right| &\leq C\,W_A(x_1), \\ |\mathbf{u}(\mathbf{x}) - \mathbf{v}_A(\mathbf{x})| &\leq C\,(W_\infty(x_1) - W_A(x_1)), \\ \left|\tfrac{\partial \mathbf{u}}{\partial x_j}(\mathbf{x}) - \tfrac{\partial \mathbf{v}_A}{\partial x_j}(\mathbf{x})\right| &\leq C\,(W_\infty(x_1) - W_A(x_1)) \end{aligned} \tag{24}$$

for $\mathbf{x} \in T_{h'}$, $j = 1, 2$, with certain generic constants C, where

$$W_A(x_1) := \int_{-A}^{A} (1 + |x_1 - y_1|)^{-3/2}\, w(y_1)\, dy_1, \qquad x_1 \in \mathbb{R}.$$

Recalling (18), we can estimate by (21)–(23) and Lemma 4.5 that

$$\begin{aligned} \int_{-A}^{A} |w(x_1)|^2\, dx_1 \leq -C \int_{S(A)} &\left\{ \mu n_2 \left| \frac{\partial \mathbf{u}}{\partial \mathbf{n}} \right|^2 + (\lambda + \mu) n_2 |\operatorname{div} \mathbf{u}|^2 \right\} ds \\ \leq C \Big\{ &|I(h', A)[\mathbf{u}] - I(h', A)[\mathbf{v}_A]| \\ &+ |I(h', A)[\mathbf{v}_A] - I(h', \infty)[\mathbf{v}_A]| \\ &+ 2k_s \Big[|K(h', \infty)[\mathbf{v}_A] - K(h', A)[\mathbf{v}_A]| \\ &\qquad + |K(h', A)[\mathbf{v}_A] - K(h', A)[\mathbf{u}]| \Big] \\ &+ |J_1(h', A)[\mathbf{u}]| + 2k_s |J_2(h', A)[\mathbf{u}]| \Big\}. \end{aligned} \tag{25}$$

From (24) there now follows, with some positive constant C,

$$\left.\begin{array}{r} |I(h', A)[\mathbf{v}_A] - I(h', \infty)[\mathbf{v}_A|, \\ |K(h', \infty)[\mathbf{v}_A] - K(h', A)[\mathbf{v}_A]| \end{array}\right\} \leq C \int_{\mathbb{R} \setminus [-A, A]} W_A^2(x_1)\, dx_1$$

and

$$\left.\begin{array}{r} |I(h', A)[\mathbf{u}] - I(h', A)[\mathbf{v}_A]|, \\ |K(h', A)[\mathbf{v}_A] - K(h', A)[\mathbf{u}]| \end{array}\right\} \leq C \int_{-A}^{A} (W_\infty(x_1) - W_A(x_1))\, W_\infty(x_1)\, dx_1,$$

so that we finally conclude, for some constant $c > 0$ and all $A > 0$,

$$\begin{aligned} \int_{-A}^{A} |w(x_1)|^2\, dx_1 \leq c \Big\{ &\int_{\mathbb{R} \setminus [-A, A]} W_A^2(x_1)\, dx_1 \\ &+ \int_{-A}^{A} (W_\infty(x_1) - W_A(x_1))\, W_\infty(x_1)\, dx_1 \\ &+ |J_1(h', A)[\mathbf{u}]| + |J_2(h', A)[\mathbf{u}]| \Big\}. \end{aligned} \tag{26}$$

Since (26) holds and we also have from Theorem A.3 that $w \in L^\infty(\mathbb{R})$, we can apply Lemma A in [7] to obtain that $w \in L^2(\mathbb{R})$ and, noting (23), that for all $A_0 > 0$,

$$(1 + \|f'\|^2_{\infty;\mathbb{R}})^{-1/2} \int_S |\mathbf{Pu}|^2 \, ds \le \int_{-\infty}^{\infty} |w(x_1)|^2 \, dx_1$$
$$\le c \sup_{A > A_0} \{|J_1(h', A)[\mathbf{u}]| + |J_2(h', A)[\mathbf{u}]\}|. \tag{27}$$

For $x \in D_{h'}$ with $|x_1| > 0$, we now deduce by Theorem 2.2 (iv), Theorem 4.3, and the Cauchy–Schwartz inequality that

$$\begin{aligned} |\mathbf{u}(\mathbf{x})|^2 &\le 2 \left\{ \int_{S \setminus S(|x_1|/2)} |\Gamma_{D,h}(\mathbf{x}, \mathbf{y}) \, \mathbf{Pu}(\mathbf{y})| \, ds(\mathbf{y}) \right\}^2 \\ &\quad + 2 \left\{ \int_{S(|x_1|/2)} |\Gamma_{D,h}(\mathbf{x}, \mathbf{y}) \, \mathbf{Pu}(\mathbf{y})| \, ds(\mathbf{y}) \right\}^2 \\ &\le C_1 \int_{S \setminus S(|x_1|/2)} |\mathbf{Pu}|^2 \, ds + C_2 \left(\frac{|x_1|}{2} \right)^{-2}, \end{aligned}$$

where

$$C_1 = 16 \sup_{\mathbf{x} \in D_{h'}} \int_S \max_{j,k=1,2} |\Gamma_{D,h,jk}(\mathbf{x}, \mathbf{y})|^2 \, ds(\mathbf{y}) < \infty$$

by Remark 2.4 and

$$C_2 = 32 \, \|\mathcal{H}\|^2_{C([0,h'-h]^2)} \, (1 + \|f'\|_{\infty;\mathbb{R}})^{1/2} \, \|\mathbf{Pu}\|_{[L^2(S)]^2}.$$

Thus, $\mathbf{u}(\mathbf{x}) \to 0$ as $|x_1| \to \infty$ ($\mathbf{x} \in D_{h'}$), uniformly in x_2. From Lemma A.1 and Theorem A.3 it now follows that $J_j(A)[\mathbf{u}] \to 0$ as $A \to \infty$ ($j = 1, 2$), and consequently, by (27), that $\mathbf{Pu} = 0$ on S. Recalling Theorem 4.3 once more, we conclude that $\mathbf{u} \equiv 0$ in Ω. We have thus shown the following theorem.

THEOREM 4.6. *Let* $\mathbf{u}$ *and* $\mathbf{v}$ *be solutions of Problem* 4.1 *with the same Dirichlet data* $\mathbf{g}$. *Then* $\mathbf{u} \equiv \mathbf{v}$ *in* Ω.

Appendix. Regularity results. The following regularity results for solutions to the Navier equation, which are special cases of general results for systems of second order elliptic equations, are used in this paper.

LEMMA A.1. *Given a domain* $G \subset \mathbb{R}^2$, *let* $\mathbf{u} \in [L^\infty(G)]^2$ *be a solution to the Navier equation* (3) *in* G *in a distributional sense. Assume* $G' \subset\subset G$ *and set* $d := d(\partial G', \partial G)$. *Then* $\mathbf{u} \in [C^1(\overline{G'})]^2$ *and, for all* $\mathbf{x} \in G'$,

$$|\operatorname{grad} u_k(\mathbf{x})| \le C \, (1 + d^{-1}) \, \|\mathbf{u}\|_{\infty;G}, \qquad k = 1, 2,$$

where C *is only dependent on* μ, λ, *and* ω.

Proof. The proof follows from application of estimates in Fichera [9] and Sobolev's imbedding theorem. □

By applications of this result we immediately obtain the following corollary.

COROLLARY A.2. *Given a domain* $G \subset \mathbb{R}^2$, *let* $(\mathbf{v}_n) \subset [L^\infty(G)]^2$ *be a sequence of solutions to the Navier equation in* G *and, for some vector field* $\mathbf{v}$, *suppose that*

$\mathbf{v}_n(\mathbf{x}) \to \mathbf{v}(\mathbf{x})$ *uniformly on compact subsets of* G. *Then* $\mathbf{v} \in [C^2(G)]^2$ *and is a solution to the Navier equation in* G.

The next result can be obtained in a manner very similar to that employed for scalar elliptic equations (see, e.g., Gilbarg and Trudinger [11]). A detailed proof is given in [3, Theorem 2.7].

THEOREM A.3. *Let* $\mathbf{u} \in [C^2(\Omega) \cap C(\bar{\Omega}) \cap H^1_{loc}(\bar{\Omega})]^2$ *be a solution to the Navier equation in* Ω, *bounded in* D_H *for some* $H > \sup f$, *with* $\mathbf{u} = 0$ *on* S. *Then* $\mathbf{u} \in \mathcal{V}^{1,\alpha}(\bar{\Omega})$ *and, for any* $H > \sup f$, $u \in C^{1,\alpha}(\overline{D_H})$ *with*

$$\|\mathbf{u}\|_{1,\alpha;D_H} \leq C \, \|\mathbf{u}\|_{\infty;D_H}, \tag{28}$$

where C *is a constant dependent only on* λ, μ, ω, H, *and* $\|f\|_{1,\alpha;\mathbb{R}}$.

REFERENCES

[1] T. ARENS, *A new integral equation formulation for the scattering of plane elastic waves by diffraction gratings*, J. Integral Equations Appl., 11 (1999), pp. 275–297.

[2] T. ARENS, *The scattering of plane elastic waves by a one-dimensional periodic surface*, Math. Methods Appl. Sci., 22 (1999), pp. 55–72.

[3] T. ARENS, *The Scattering of Elastic Waves by Rough Surfaces*, Ph.D. thesis, Brunel University, Uxbridge, UK, 2000.

[4] S. N. CHANDLER-WILDE, *The impedance boundary value problem for the Helmholtz equation in a half-plane*, Math. Methods Appl. Sci., 20 (1997), pp. 813–840.

[5] S. N. CHANDLER-WILDE, C. R. ROSS, AND B. ZHANG, *Scattering by rough surfaces*, in Proceedings of the Fourth International Conference on Mathematical and Numerical Aspects of Wave Propagation, J. DeSanto, ed., SIAM, Philadelphia, 1998, pp. 164–168.

[6] S. N. CHANDLER-WILDE AND B. ZHANG, *Electromagnetic scattering by an inhomogenous conducting or dielectric layer on a perfectly conducting plate*, Proc. Roy. Soc. London Ser. A, 454 (1998), pp. 519–542.

[7] S. N. CHANDLER-WILDE AND B. ZHANG, *A uniqueness result for scattering by infinite rough surfaces*, SIAM J. Appl. Math., 58 (1998), pp. 1774–1790.

[8] S. N. CHANDLER-WILDE AND B. ZHANG, *Scattering of electromagnetic waves by rough interfaces and inhomogeneous layers*, SIAM J. Math. Anal., 30 (1999), pp. 559–583.

[9] G. FICHERA, *Linear Elliptic Differential Systems and Eigenvalue Problems*, Lecture Notes in Math. 8, Springer, Berlin, 1965.

[10] J. T. FOKKEMA AND P. M. VAN DEN BERG, *Elastodynamic diffraction by a periodic rough surface*, J. Acoust. Soc. Amer., 62 (1977), pp. 1095–1101.

[11] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer, Berlin, 1983.

[12] R. KRESS, *Inverse elastic scattering from a crack*, Inverse Problems, 12 (1996), pp. 667–684.

[13] V. D. KUPRADZE, *Potential Methods in the Theory of Elasticity*, Israeli Program for Scientific Translations, Jerusalem, 1965.

[14] A. LAKHTAKIA, V. K. VARADAN, V. V. VARADAN, AND D. J. N. WALL, *The T-matrix approach for scattering by a traction-free periodic rough surface*, J. Acoust. Soc. Amer., 76 (1984), pp. 1839–1846.

[15] A. LAKHTAKIA, V. V. VARADAN, AND V. K. VARADAN, *Reflection characteristics of an elastic slab containing a periodic array of circular elastic cylinders: P and SV wave analysis*, J. Acoust. Soc. Amer., 83 (1988), pp. 1267–1275.

[16] B. ZHANG AND S. N. CHANDLER-WILDE, *Acoustic scattering by an inhomogeneous layer on a rigid plate*, SIAM J. Appl. Math., 58 (1998), pp. 1931–1950.

SIAM J. MATH. ANAL.
Vol. 33, No. 2, pp. 477–488

GLOBAL EXISTENCE FOR SYSTEMS OF NONLINEAR WAVE EQUATIONS IN 3D WITH MULTIPLE SPEEDS*

THOMAS C. SIDERIS† AND SHU-YI TU‡

Abstract. Global smooth solutions to the initial value problem for systems of nonlinear wave equations with multiple propagation speeds will be constructed in the case of small initial data and nonlinearities satisfying the null condition.

Key words. systems of nonlinear wave equations, global existence, null condition

AMS subject classification. 35L70

PII. S0036141000378966

1. Introduction.

This paper is concerned with the Cauchy problem for coupled systems of quasi-linear wave equations in three space dimensions of the form

$$\partial_t^2 u^k - c_k^2 \triangle u^k = C^{jk}_{\alpha\beta}(\partial u)\partial_\alpha\partial_\beta u^j, \quad k = 1, \ldots, m,$$

subject to suitably small initial conditions. We assume that the propagation speeds are distinct, and we refer to this situation as the nonrelativistic case. Here, ∂u stands for the full space-time gradient, and $C^{jk}_{\alpha\beta}(\xi) = O(|\xi|)$ are smooth functions near the origin in $\mathbb{R}^{4m}$. We shall construct a unique global classical solution, provided that the coefficients of the nonlinear terms satisfy the null condition, which permits only certain special nonlinear self-interactions of the kth component of the solution in the kth equation. This nonrelativistic system serves as a simplified model for wave propagation problems with different speeds, such as nonlinear elasticity, charged plasmas, and magneto-hydrodynamics.

The main difficulty in the nonrelativistic case is that the smaller symmetry group of the linear operator weakens the form of the invariant Klainerman inequality; see section 6. In order to obtain a viable $L^\infty - L^2$ estimate for solutions, we utilize an additional set of weighted L^2 estimates, as has been developed in [15], [19], [20]. The advantage of this method is the total avoidance of direct estimation of the fundamental solution for the linear problem as well as any type of asymptotic constructions. We treat nondivergence form nonlinearities which may contain both spatial and temporal derivatives.

In the three-dimensional (3D) relativistic (scalar) case, the null condition was first identified and shown to lead to global existence of small solutions by Christodoulou [3] and Klainerman [13]. Without it, small solutions remain smooth "almost globally" [8], [9], [12], but as examples show, arbitrarily small initial conditions can develop singularities in finite time [7], [18]. Small solutions always exist globally in higher dimensions [11], [17], [12]. The two-dimensional (2D) relativistic case is rather more complicated. The sharpest results are given in [1], [2], but other work appeared previously in [4], [10].

*Received by the editors October 2, 2000; accepted for publication (in revised form) March 9, 2001; published electronically July 31, 2001.
http://www.siam.org/journals/sima/33-2/37896.html

†Department of Mathematics, University of California, Santa Barbara, CA 93106 (sideris@math.ucsb.edu). This author was supported in part by the National Science Foundation.

‡Department of Mathematics, St. Cloud State University, St. Cloud, MN 56301 (stu@stcloudstate.edu).

The case of nonrelativistic systems in 3D has also recently been considered by Yokoyama [21]. Under the same null condition as described below, Yokoyama establishes the existence of global small solutions. However, instead of expressing the smallness condition for the initial data in terms of a neighborhood of the origin in a Sobolev space, as we do here, Yokoyama considers only data of the form $(u, \partial_t u)|_{t=0} = \varepsilon(u_0, u_1)$ for fixed C^∞ functions with compact support. Another significant difference is that Yokoyama obtains decay of solutions through direct $L^\infty - L^\infty$ estimation of the fundamental solution for the linear wave equation. By avoiding such direct estimations, our $L^\infty - L^2$ approach is much simpler. Moreover, our estimates are sharper insofar as they do not require the logarithmic growth factors used in Proposition 3.1 in [21].

An early result for 3D nonrelativistic systems was obtained by Kovalyov [16] in the semilinear case under a strong *nonresonance condition* that ruled out all nonlinear self-interactions. The 2D case has been examined in [6] and [5] using an approach similar to [21].

The statement of the main result is given in section 3 after a summary of some standard notation. The rest of the paper presents the proof. To simplify the exposition, we truncate the nonlinearity at the quadratic level, but this entails no loss of generality since the higher-order terms do not affect the global behavior of small solutions [12].

2. Notation. Points in $\mathbb{R}^4$ will be denoted by $X = (x^0, x^1, x^2, x^3) = (t, x)$. Partial derivatives will be written as $\partial_k = \partial/\partial x^k$, $k = 0, \dots, 3$, with the abbreviations $\partial = (\partial_0, \partial_1, \partial_2, \partial_3) = (\partial_t, \nabla)$. The angular-momentum operators are defined as

$$\Omega = (\Omega_1, \Omega_2, \Omega_3) = x \wedge \nabla,$$

where $\wedge$ denotes the usual vector cross product in $\mathbb{R}^3$, and the scaling operator is defined by

$$S = t\partial_t + r\partial_r = x^\alpha \partial_\alpha. \tag{2.1}$$

The collection of these seven vector fields will be labeled as

$$\Gamma = (\Gamma_0, \dots, \Gamma_7) = (\partial, \Omega, S).$$

Instead of the usual multi-index notation, we will write $a = (a_1, \dots, a_\kappa)$ for a sequence of indices $a_i \in \{0, \dots, 7\}$ of length $|a| = \kappa$, and

$$\Gamma^a = \Gamma_{a_\kappa} \cdots \Gamma_{a_1}.$$

Suppose that b and c are disjoint subsequences of a. Then we will say $b + c = a$ if $|b| + |c| = |a|$, and $b + c < a$ if $|b| + |c| < |a|$.

The d'Alembertian will be used to denote the operator

$$\Box = \mathrm{Diag}(\Box_1, \dots, \Box_m) \quad \text{with} \quad \Box_k = \partial_t^2 - c_k^2 \triangle.$$

For convenience, we will assume that the speeds are distinct

$$c_1 > \cdots > c_m > 0.$$

It is also possible to treat the case where some of the speeds are the same; see the remark following the statement of Theorem 3.1. This operator acts on vector functions $u : \mathbb{R}^4 \to \mathbb{R}^m$. The standard energy is then defined as

$$E_1(u(t)) = \sum_{k=1}^m \int_{\mathbb{R}^3} [\, |\partial_t u^k(t,x)|^2 + c_k^2 \, |\nabla u^k(t,x)|^2 \,] \, dx,$$

and higher-order derivatives will be estimated through

$$E_\kappa(u(t)) = \sum_{|a|\le\kappa-1} E_1(\Gamma^a u(t)), \qquad \kappa = 2, 3, \ldots. \tag{2.2a}$$

In order to describe the solution space, we introduce the time-independent vector fields $\Lambda = (\Lambda_1, \ldots, \Lambda_7) = (\nabla, \Omega, r\partial_r)$. Define

$$H^\kappa_\Lambda(\mathbb{R}^3) = \{f \in L^2(\mathbb{R}^3; \mathbb{R}^m) : \Lambda^a f \in L^2,\ |a| \le \kappa\}$$

with the norm

$$\|f\|_{H^\kappa_\Lambda} = \sum_{|a|\le\kappa} \|\Lambda^a f\|_{L^2}. \tag{2.2b}$$

Solutions will be constructed in the space $\dot{H}^\kappa_\Gamma(T)$ obtained by closing the set $C^\infty([0,T); C^\infty_0(\mathbb{R}^3, \mathbb{R}^m))$ in the norm $\sup_{0\le t<T} E^{1/2}_\kappa(u(t))$. Thus,

$$\dot{H}^\kappa_\Gamma(T) \subset \left\{ u(t,x) : \partial u(t,\cdot) \in \bigcap_{j=0}^{\kappa-1} C^j([0,T); H^{\kappa-1-j}_\Lambda) \right\}.$$

By (6.1) it will follow that $\dot{H}^\kappa_\Gamma(T) \subset C^{\kappa-2}([0,T) \times \mathbb{R}^3; \mathbb{R}^m)$.

An important intermediate role will be played by the weighted norm

$$\mathcal{X}_\kappa(u(t)) = \sum_{k=1}^m \sum_{|a|=2} \sum_{|b|\le\kappa-2} \|\langle c_k t - |x|\rangle \partial^a \Gamma^b u^k(t)\|_{L^2(\mathbb{R}^3)}, \tag{2.2c}$$

where we use the notation $\langle\rho\rangle = (1 + |\rho|^2)^{1/2}$.

3. Main result. Consider the initial value problem for a coupled nonlinear system of the form

$$\Box u = N(u,u) \tag{3.1}$$

in which the components of the quadratic nonlinearity depend on the form

$$N^k(u,v) = C^{ijk}_{\alpha\beta\gamma} \partial_\alpha u^i \partial_\beta \partial_\gamma v^j. \tag{3.2a}$$

Summation is performed over repeated indices regardless of their position, up or down. Greek indices range from 0 to 3 and Latin indices from 1 to m.

Existence of solutions depends on the energy method which requires the system to be symmetric:

$$C^{ijk}_{\alpha\beta\gamma} = C^{ikj}_{\alpha\beta\gamma} = C^{ijk}_{\alpha\gamma\beta}. \tag{3.2b}$$

The key assumption necessary for global existence is the following *null* condition which says that the self-interaction of each wave family is nonresonant:

$$C^{kkk}_{\alpha\beta\gamma} X_\alpha X_\beta X_\gamma = 0 \quad \text{for all} \quad X \in \mathcal{N}_k, \quad k = 1, \ldots, m, \tag{3.2c}$$

with the null cones

$$\mathcal{N}_k = \{X \in \mathbb{R}^4 : x_0^2 - c_k^2(x_1^2 + x_2^2 + x_3^2) = 0\}.$$

Theorem 3.1. *Assume that the nonlinear terms in* (3.2a) *satisfy the symmetry and null conditions* (3.2b), (3.2c). *Then the initial value problem for* (3.1) *with initial data*

$$\partial_\alpha u(0) \in H^{\kappa-1}_\Lambda(\mathbb{R}^3), \quad \kappa \geq 9,$$

satisfying

$$E^{1/2}_{\kappa-2}(u(0)) \ \exp\ C E^{1/2}_{\kappa}(u(0)) < \varepsilon, \tag{3.3}$$

with ε sufficiently small, has a unique global solution $u \in \dot{H}^\kappa_\Gamma(T)$ for every $T > 0$. The solution satisfies the bounds

$$E^{1/2}_{\kappa-2}(u(t)) < 2\varepsilon \quad \text{and} \quad E_\kappa(u(t)) \leq 4E_\kappa(u(0))\langle t\rangle^{C\varepsilon}.$$

Remark. We briefly discuss the case when some of the speeds are repeated. Suppose that only $\ell < m$ of the speeds $c_1 = c_{k_1} > c_{k_2} > \cdots > c_{k_\ell}$ are distinct. For $p = 1, \ldots, \ell$, let $I_p = \{k : 1 \leq k \leq m,\ c_k = c_{k_p}\}$. The null condition is now extended to be

$$C^{ijk}_{\alpha\beta\gamma} X_\alpha X_\beta X_\gamma = 0 \quad \text{for all} \quad X \in \mathcal{N}_{k_p},\ (i,j,k) \in I^3_p,\ p = 1, \ldots, \ell.$$

The proof can easily be adjusted to handle this more general case.

4. Commutation and null forms. In preparation for the energy estimates, we need to consider the commutation properties of the vector fields Γ with respect to the nonlinear terms. It is necessary to verify that the null structure is preserved upon differentiation.

Lemma 4.1. *Let u be solution u of* (3.1) *in $\dot{H}^\kappa_\Gamma(T)$. Assume that the null condition* (3.2c) *holds for the nonlinearity in* (3.2a). *Then for $|a| \leq \kappa - 1$,*

$$\Box\Gamma^a u = \sum_{b+c+d=a} N_d(\Gamma^b u, \Gamma^c u)$$

in which each N_d is a quadratic nonlinearity of the form (3.2a) *satisfying* (3.2c). *Moreover, if $b + c = a$, then $N_d = N$.*

Proof. First we note the well-known facts that

$$[\partial, \Box] = 0, \quad [\Omega, \Box] = 0, \quad [S, \Box] = -2\Box.$$

Recalling the definition (3.2a), we set

$$[\Gamma, N](u, v) = \Gamma N(u, v) - N(\Gamma u, v) - N(u, \Gamma v).$$

This is a quadratic nonlinearity of the form (3.2a). Thus, if $[\Gamma, N]$ is null for each Γ, then the result follows by induction. In fact, if $d = (d_1, \ldots, d_k)$, then N_d is the k-fold commutator $N_d = [\Gamma_{d_k}, [\ldots, [\Gamma_{d_1}, N]]]$.

A simple calculation shows that

$$[\partial, N](u, v) = 0 \quad \text{and} \quad [S, N](u, v) = -3N(u, v).$$

Thus, these commutators are null if N is null.

We can express the angular momentum operators as $\Omega_\lambda = \varepsilon_{\lambda\mu\nu} x_\mu \partial_\nu$, $\lambda = 1, 2, 3$, where $\varepsilon_{\lambda\mu\nu}$ is the tensor with value $+1$, -1 if $\lambda\mu\nu$ is an even, respectively, odd, permutation of 123, and with value 0 otherwise. Using this, we find that the kth component of $[\Omega_\lambda, N]$ is

$$[\Omega_\lambda, N]^k(u,v) = \widetilde{C}^{ijk}_{\alpha\beta\gamma}\partial_\alpha u^j \partial_\beta\partial_\gamma v^k$$

with

$$\widetilde{C}^{ijk}_{\alpha\beta\gamma} = [C^{ijk}_{\alpha\beta\nu}\varepsilon_{\lambda\gamma\nu} + C^{ijk}_{\nu\beta\gamma}\varepsilon_{\lambda\alpha\nu} + C^{ijk}_{\alpha\nu\gamma}\varepsilon_{\lambda\beta\nu}].$$

To see that this commutator is also null, write

$$h^k(X) = C^{kkk}_{\alpha\beta\gamma}X_\alpha X_\beta X_\gamma \quad \text{and} \quad \tilde{h}^k(X) = \widetilde{C}^{kkk}_{\alpha\beta\gamma}X_\alpha X_\beta X_\gamma.$$

Then $\tilde{h}^k(X) = -Dh^k(X)Y^\lambda$ with $Y^\lambda_\mu = \varepsilon_{\lambda\mu\nu}X_\nu$. Now the null condition says that $h^k(X) = 0$ for $X \in \mathcal{N}_k$. But since Y^λ is tangent to $\mathcal{N}_k$ at X, we have $\tilde{h}^k(X) = 0$ for $X \in \mathcal{N}_k$. This implies that $[\Omega_\lambda, N]$ is null. □

5. Estimates for null forms. The utility of the null condition is captured in the next lemma. The presence of the terms with the weight $\langle c_k t - r\rangle$ in these inequalities is explained by the absence of the Lorentz rotations in our list of vector fields Γ.

LEMMA 5.1. *Suppose that the nonlinear form $N(u,v)$ defined in* (3.2a) *satisfies the null condition* (3.2c). *Set $c_0 = \min\{c_k/2 : k = 1,\dots,m\}$. For u, v, $w \in C^2([0,T]\times\mathbb{R}^3;\mathbb{R}^m)$, and $r \geq c_0 t$, we have at any point $X = (t,x)$*

$$\text{(5.1a)} \qquad |C^{kkk}_{\alpha\beta\gamma}\partial_\alpha u^k\partial_\beta\partial_\gamma v^k| \leq \frac{C}{\langle X\rangle}\Big[|\Gamma u^k||\partial^2 v^k| + |\partial u^k||\partial\Gamma v^k| + \langle c_k t - r\rangle|\partial u^k||\partial^2 v^k|\Big]$$

and

$$\text{(5.1b)} \; |C^{kkk}_{\alpha\beta\gamma}\partial_\alpha u^k\partial_\beta v^k\partial_\gamma w^k| \leq \frac{C}{\langle X\rangle}\Big[|\Gamma u^k||\partial v^k||\partial w^k| + |\partial u^k||\Gamma v^k||\partial w^k| + |\partial u^k||\partial v^k||\Gamma w^k| + \langle c_k t - r\rangle|\partial u^k||\partial v^k||\partial w^k|\Big]$$

in which $\langle X\rangle = (1+|X|^2)^{1/2}$.

Proof. Spatial derivatives have the decomposition

$$\nabla = \frac{x}{r}\partial_r - \frac{x}{r^2}\wedge\Omega.$$

So if we introduce the two operators $D^\pm_k = \frac{1}{2}(\partial_t \pm c_k\partial_r)$ and the null vectors $Y^\pm_k = (1, \pm x/c_k r) \in \mathcal{N}_k$, we obtain

$$\text{(5.2)} \qquad (\partial_t, \nabla) = (Y^-_k D^-_k + Y^+_k D^+_k) - \Big(0, \frac{x}{r^2}\wedge\Omega\Big).$$

On the other hand, if we write

$$D^+_k = \frac{c_k}{c_k t + r}S - \frac{c_k t - r}{c_k t + r}D^-_k,$$

the formula (5.2) can be transformed into

$$\partial = Y^-_k D^-_k - \frac{c_k t - r}{c_k t + r}Y^+_k D^-_k + \frac{c_k}{c_k t + r}Y^+_k S - \Big(0, \frac{x}{r^2}\wedge\Omega\Big).$$

Thus, we have

$$\partial \equiv Y_k^- D_k^- + R. \tag{5.3a}$$

Now, we may assume that $|X| \geq 1$, for otherwise the estimates are trivial. But then it follows that $1/r$ and $1/(c_k t + r)$ are bounded by $C/\langle X \rangle$, and as a consequence we have

$$|Ru| \leq C\langle X\rangle^{-1}[|\Gamma u| + \langle c_k t - r\rangle |\partial u|]. \tag{5.3b}$$

Using (5.3a), we have

$$\begin{aligned} C^{kkk}_{\alpha\beta\gamma}\partial_\alpha u^k \partial_\beta\partial_\gamma v^k = C^{kkk}_{\alpha\beta\gamma}[&Y^-_{k\alpha}Y^-_{k\beta}Y^-_{k\gamma}D^-_k u^k (D^-_k)^2 v^k + R_\alpha u^k \partial_\beta\partial_\gamma v^k \\ &+ Y^-_{k\alpha}D^-_k u^k R_\beta \partial_\gamma v^k + Y^-_{k\alpha}D^-_k u^k Y^-_{k\beta} D^-_k R_\gamma v^k]. \end{aligned} \tag{5.4}$$

The first term in (5.4) vanishes since N obeys the null condition, and by (5.3b) the remaining terms in (5.4) have the estimate (5.1a).

The proof of (5.1b) is similar. □

6. Sobolev inequalities. The following Sobolev inequalities involve only the angular momentum operators since we are in the nonrelativistic case. The weight $\langle ct - r\rangle$ compensates for this. We use the notation defined in (2.2a), (2.2b), (2.2c).

Lemma 6.1. *Let $u \in \dot H^\kappa_\Gamma(T)$ with $\mathcal{X}_\kappa(u(t)) < \infty$.*

$$\langle r\rangle^{1/2}|\Gamma^a u(t,x)| \leq C E^{1/2}_\kappa(u(t)), \quad |a| + 2 \leq \kappa, \tag{6.1}$$

$$\langle r\rangle|\partial\Gamma^a u(t,x)| \leq C E^{1/2}_\kappa(u(t)), \quad |a| + 3 \leq \kappa, \tag{6.2}$$

$$\langle r\rangle\langle c_i t - r\rangle^{1/2}|\partial\Gamma^a u^i(t,x)| \leq C\Big[E^{1/2}_\kappa(u(t)) + \mathcal{X}_\kappa(u(t))\Big], \quad |a| + 3 \leq \kappa, \tag{6.3}$$

$$\langle r\rangle\langle c_i t - r\rangle|\partial^2\Gamma^a u^i(t,x)| \leq C\mathcal{X}_\kappa(u(t)), \quad |a| + 4 \leq \kappa. \tag{6.4}$$

Proof. This result is essentially Proposition 3.3 in [20] (see also [14]). □

7. Weighted decay estimates. The main extra step in the nonrelativistic case is to control the weighted norm $\mathcal{X}_\kappa(u(t))$. This will be accomplished in this section by a type of bootstrap argument.

Lemma 7.1. *Let $u \in \dot H^\kappa_\Gamma(T)$. Then*

$$\mathcal{X}_\kappa(u(t)) \leq C\left[E^{1/2}_\kappa(u(t)) + \sum_{|a|\leq\kappa-2}\|(t+r)\,\Box\Gamma^a u(t)\|_{L^2}\right]. \tag{7.1}$$

Proof. Recall that the weighted norm involves derivatives in the form $\partial^2\Gamma^a u$. In the case when $\partial^2 = \nabla\partial$, the result was given in Lemma 3.1 of [15]. Otherwise, if $\partial^2 = \partial_t^2$, then the result is an immediate consequence of (2.10) in [15]. □

Now we assume that u solves the nonlinear PDE.

Lemma 7.2. *Let $u \in \dot H^\kappa_\Gamma(T)$ be a solution of* (3.1)*. Define $\kappa' = \left[\frac{\kappa-1}{2}\right] + 3$. Then for all $|a| \leq \kappa - 2$,*

$$\|(t+r)\Box\Gamma^a u(t)\|_{L^2} \leq C[\mathcal{X}_{\kappa'}(u(t))E^{1/2}_\kappa(u(t)) + \mathcal{X}_\kappa(u(t))E^{1/2}_{\kappa'}(u(t))]. \tag{7.2}$$

Proof. By Lemma 4.1, we must estimate terms of the form

$$\|(t+r)\partial\Gamma^b u^i \partial^2\Gamma^c u^j\|_{L^2},$$

but since $(t+r) \leq C\langle r\rangle\langle c_j t - r\rangle$, we will consider

$$\|\langle r\rangle\langle c_j t - r\rangle \partial\Gamma^b u^i \partial^2\Gamma^c u^j\|_{L^2} \tag{7.3}$$

with $b+c \leq a$ and $|a| \leq \kappa - 2$.

Let $m = \left[\frac{\kappa-1}{2}\right] = \kappa' - 3$. We separate two cases: either $|b| \leq m$ or $|c| \leq m-1$. In the first case, (7.3) is estimated as follows using (6.2):

$$\|\langle r\rangle \partial\Gamma^b u^i\|_{L^\infty} \|\langle c_j t - r\rangle \partial^2\Gamma^c u^j\|_{L^2} \leq C E_{\kappa'}^{1/2}(u(t))\mathcal{X}_\kappa(u(t)).$$

Otherwise, we use (6.4) to estimate (7.3) by

$$\|\partial\Gamma^b u^i\|_{L^2} \|\langle r\rangle\langle c_j t - r\rangle \partial^2\Gamma^c u^j\|_{L^\infty} \leq C E_\kappa^{1/2}(u(t))\mathcal{X}_{\kappa'}(u(t)). \qquad \square$$

The next result gains control of the weighted norm by the energy. We distinguish two different energies, the smaller of which must remain small. In the next section, we will allow the larger energy to grow polynomially in time.

LEMMA 7.3. *Let $u \in \dot{H}^\kappa_\Gamma(T)$, $\kappa \geq 8$, be a solution of* (3.1)*. Define $\mu = \kappa - 2$, and assume that*

$$\varepsilon_0 \equiv \sup_{0 \leq t < T} E_\mu^{1/2}(u(t))$$

is sufficiently small. Then for $0 \leq t < T$,

$$\mathcal{X}_\mu(u(t)) \leq C E_\mu^{1/2}(u(t)) \tag{7.4a}$$

and

$$\mathcal{X}_\kappa(u(t)) \leq C E_\kappa^{1/2}(u(t)). \tag{7.4b}$$

Proof. Let $\mu' = \left[\frac{\mu-1}{2}\right] + 3$, $\mu = \kappa - 2$. Since $\mu \geq 6$, we have $\mu' \leq \mu$. Thus, by Lemmas 7.1 and 7.2, we find using our assumption that

$$\mathcal{X}_\mu(u(t)) \leq C[E_\mu^{1/2}(u(t)) + \varepsilon_0 \mathcal{X}_\mu(u(t))].$$

Thus, if ε_0 is small enough, the bound (7.4a) results.

Again, since $\kappa \geq 8$, we have $\kappa' = \left[\frac{\kappa-1}{2}\right] + 3 \leq \mu = \kappa - 2$. From Lemmas 7.1 and 7.2 we now have

$$\mathcal{X}_\kappa(u(t)) \leq C[E_\kappa^{1/2}(u(t)) + \mathcal{X}_\mu(u(t))E_\kappa^{1/2}(u(t)) + \mathcal{X}_\kappa(u(t))E_\mu^{1/2}(u(t))].$$

If we apply (7.4a) and our assumption, then

$$\mathcal{X}_\kappa(u(t)) \leq C[E_\kappa^{1/2}(u(t)) + \varepsilon_0 \mathcal{X}_\kappa(u(t))],$$

from which (7.4b) follows. $\square$

8. Energy estimates.

General energy method. In this section we shall complete the proof of Theorem 3.1. Assume that $u(t) \in \dot{H}^\kappa_\Gamma(T)$ is a local solution of the initial value problem for (3.1). Our task will be to show that $E_\kappa(u(t))$ remains finite for all $t \geq 0$. To do so, we will derive a pair of coupled differential inequalities for (modifications of) $E_\kappa(u(t))$ and $E_\mu(u(t))$ with $\mu = \kappa - 2$. If (3.3) holds, then $E_\mu^{1/2}(u(0)) < \varepsilon$. Suppose that T_0 is the largest time such that $E_\mu^{1/2}(u(t)) < 2\varepsilon$ for $0 \leq t < T_0$ with ε small enough so that Lemma 7.3 is valid. All of the following computations will be valid on this time interval.

Following the energy method, we have for any $\nu = 1, \dots, \kappa$,

$$E_\nu'(u(t)) = \sum_{|a|\leq \nu-1} \int \langle \Box \Gamma^a u(t), \partial_t \Gamma^a u(t) \rangle dx,$$

and from Lemma 4.1, this takes the form

$$E_\nu'(u(t)) = \sum_{|a|\leq \nu-1} \sum_{b+c+d=a} \int \langle N_d(\Gamma^b u, \Gamma^c u), \partial_t \Gamma^a u \rangle dx. \tag{8.1}$$

Terms in (8.1) with $b = 0$, $c = a$, and $|a| = \nu - 1$ are handled with the aid of the symmetry condition (3.2b) which allows us to integrate by parts as follows. Recall that from Lemma 4.1, $N_d = N$ when $b + c = a$.

$$\begin{aligned}
\int \langle N(u, \Gamma^a u), \partial_t \Gamma^a u \rangle dx &= C^{ijk}_{\alpha\beta\gamma} \int \partial_\alpha u^i \partial_\beta \partial_\gamma \Gamma^a u^j \partial_t \Gamma^a u^k dx \\
&= C^{ijk}_{\alpha\beta\gamma} \int \partial_\gamma [\partial_\alpha u^i \partial_\beta \Gamma^a u^j \partial_t \Gamma^a u^k] dx \\
&\quad - C^{ijk}_{\alpha\beta\gamma} \int \partial_\alpha \partial_\gamma u^i \partial_\beta \Gamma^a u^j \partial_t \Gamma^a u^k dx \\
&\quad - C^{ijk}_{\alpha\beta\gamma} \int \partial_\alpha u^i \partial_\beta \Gamma^a u^j \partial_t \partial_\gamma \Gamma^a u^k dx \\
&= C^{ijk}_{\alpha\beta 0} \partial_t \int \partial_\alpha u^i \partial_\beta \Gamma^a u^j \partial_t \Gamma^a u^k dx \\
&\quad - C^{ijk}_{\alpha\beta\gamma} \int \partial_\alpha \partial_\gamma u^i \partial_\beta \Gamma^a u^j \partial_t \Gamma^a u^k dx \\
&\quad - \frac{1}{2} C^{ijk}_{\alpha\beta\gamma} \int \partial_\alpha u^i \partial_t [\partial_\beta \Gamma^a u^j \partial_\gamma \Gamma^a u^k] dx \\
&= \frac{1}{2} C^{ijk}_{\alpha\beta\gamma} \eta_{\gamma\delta}\ \partial_t \int \partial_\alpha u^i \partial_\beta \Gamma^a u^j \partial_\delta \Gamma^a u^k dx \\
&\quad - C^{ijk}_{\alpha\beta\gamma} \int \partial_\alpha \partial_\gamma u^i \partial_\beta \Gamma^a u^j \partial_t \Gamma^a u^k dx \\
&\quad + \frac{1}{2} C^{ijk}_{\alpha\beta\gamma} \int \partial_t \partial_\alpha u^i \partial_\beta \Gamma^a u^j \partial_\gamma \Gamma^a u^k dx,
\end{aligned}$$

using the symbol $\eta_{\gamma\delta} = \mathrm{Diag}[1, -1, -1, -1]$. The first term above can be absorbed into the energy as a lower order perturbation. Define

$$\widetilde{E}_\nu(u(t)) = E_\nu(u(t)) - \frac{1}{2} \sum_{|a|=\nu-1} C^{ijk}_{\alpha\beta\gamma} \eta_{\gamma\delta} \int \partial_\alpha u^i \partial_\beta \Gamma^a u^j \partial_\delta \Gamma^a u^k dx.$$

The perturbation is bounded by $C\|\partial u\|_{L^\infty} E_\nu(u(t))$, but by (6.2), the maximum norm $\|\partial u\|_{L^\infty}$ is controlled by $E_3^{1/2}(u(t)) \leq E_\mu^{1/2}(u(t)) < 2\varepsilon$. Thus, for small solutions we have

$$(1/2)E_\nu(u(t)) \leq \widetilde{E}_\nu(u(t)) \leq 2E_\nu(u(t)). \tag{8.2}$$

Returning to (8.1), we have derived the energy identity

$$\begin{aligned}\widetilde{E}'_\nu(u(t)) = &\sum_{|a|\leq \nu-1} \sum_{\substack{b+c+d=a \\ |a|\neq \nu-1}} \int \langle N_d(\Gamma^b u, \Gamma^c u), \partial_t \Gamma^a u\rangle dx \\ &+ \sum_{|a|=\nu-1} \Bigg[\sum_{\substack{b+c=a \\ c\neq a}} \int \langle N(\Gamma^b u, \Gamma^c u), \partial_t \Gamma^a u\rangle dx. \\ &\qquad - C^{ijk}_{\alpha\beta\gamma} \int \partial_\alpha \partial_\gamma u^i \partial_\beta \Gamma^a u^j \partial_t \Gamma^a u^k dx \\ &\qquad + \frac{1}{2} C^{ijk}_{\alpha\beta\gamma} \int \partial_t \partial_\alpha u^i \partial_\beta \Gamma^a u^j \partial_\gamma \Gamma^a u^k dx \Bigg].\end{aligned} \tag{8.3}$$

Higher energy. For the first series of estimates we take $\nu = \kappa$ in (8.3). We immediately obtain

$$\widetilde{E}'_\kappa(u(t)) \leq C \sum_{i,j,k} \sum_{|a|\leq \kappa-1} \sum_{\substack{b+c\leq a \\ |c|\leq \kappa-2}} \|\partial \Gamma^b u^i \partial^2 \Gamma^c u^j\|_{L^2} \|\partial \Gamma^a u^k\|_{L^2}. \tag{8.4}$$

In some cases, the indices i and j have been interchanged. In the sum on the right-hand side of (8.4), we have either $|b| \leq \kappa'$ or $|c| \leq \kappa' - 1$ with $\kappa' = \left[\frac{\kappa}{2}\right]$. Note that since $\kappa \geq 9$, we have $\kappa' + 3 \leq \kappa - 2 = \mu$. We will also use that $\langle t\rangle \leq C\langle r\rangle \langle c_j t - r\rangle$.

In the first case, we estimate using (6.2) and (7.4b) as follows:

$$\begin{aligned}\|\partial \Gamma^b u^i \partial^2 \Gamma^c u^j\|_{L^2} &\leq C\langle t\rangle^{-1} \|\langle r\rangle \partial \Gamma^b u^i\|_{L^\infty} \|\langle c_j t - r\rangle \partial^2 \Gamma^c u^j\|_{L^2} \\ &\leq C\langle t\rangle^{-1} E^{1/2}_{|b|+3}(u(t)) \mathcal{X}_\kappa(u(t)) \\ &\leq C\langle t\rangle^{-1} E^{1/2}_\mu(u(t)) E^{1/2}_\kappa(u(t)).\end{aligned}$$

In the second case, we use (6.4) and then (7.4a):

$$\begin{aligned}\|\partial \Gamma^b u^i \partial^2 \Gamma^c u^j\|_{L^2} &\leq C\langle t\rangle^{-1} \|\partial \Gamma^b u^i\|_{L^2} \|\langle r\rangle \langle c_j t - r\rangle \partial^2 \Gamma^c u^j\|_{L^\infty} \\ &\leq C\langle t\rangle^{-1} E^{1/2}_\kappa(u(t)) \mathcal{X}_{|c|+4}(u(t)) \\ &\leq C\langle t\rangle^{-1} E^{1/2}_\kappa(u(t)) \mathcal{X}_\mu(u(t)) \\ &\leq C\langle t\rangle^{-1} E^{1/2}_\kappa(u(t)) E^{1/2}_\mu(u(t)).\end{aligned}$$

Going back to (8.4) and recalling (8.2), we have established the inequality

$$\begin{aligned}\widetilde{E}'_\kappa(u(t)) &\leq C\langle t\rangle^{-1} E^{1/2}_\mu(u(t)) E_\kappa(u(t)) \\ &\leq C\langle t\rangle^{-1} \widetilde{E}^{1/2}_\mu(u(t)) \widetilde{E}_\kappa(u(t)).\end{aligned} \tag{8.5}$$

Lower energy. The second series of energy estimates will exploit the null condition. We return to (8.3) now with $\nu = \mu = \kappa - 2$. The resulting integrals on the right-hand side of (8.3) will be subdivided into separate integrals over the regions $r \leq c_0 t$ and $r \geq c_0 t$. Recall that the constant c_0 was defined in Lemma 5.1.

Inside the cones. On the region $r \leq c_0 t$, we have that the right-hand side of (8.3) is bounded above by

$$\sum_{i,j,k} \sum_{|a| \leq \mu - 1} \sum_{\substack{b+c \leq a \\ |c| \leq \mu - 2}} \|\partial \Gamma^b u^i \partial^2 \Gamma^c u^j \partial \Gamma^a u^k\|_{L^1(r \leq c_0 t)}.$$

Since $r \leq c_0 t$, we have that $\langle c_i t - r\rangle \geq C\langle t\rangle$ for each $i = 1, \ldots, m$. Thus, using (6.3), a typical term can be estimated by

$$\begin{aligned} C\langle t\rangle^{-3/2} &\|\langle c_i t - r\rangle^{1/2} \partial \Gamma^b u^i \langle c_j t - r\rangle \partial^2 \Gamma^c u^j \partial \Gamma^a u^k\|_{L^1(r \leq c_0 t)} \\ &\leq C\langle t\rangle^{-3/2} \|\langle c_i t - r\rangle^{1/2} \partial \Gamma^b u^i\|_{L^\infty} \|\langle c_j t - r\rangle \partial^2 \Gamma^c u^j\|_{L^2} \|\partial \Gamma^a u^k\|_{L^2} \\ &\leq C\langle t\rangle^{-3/2} \Big[E^{1/2}_{|b|+3}(u(t)) + \mathcal{X}_{|b|+3}(u(t))\Big] \mathcal{X}_{|c|+2}(u(t)) E^{1/2}_{\mu}(u(t)). \end{aligned}$$

In the preceding, we have $|b| + 3 \leq \kappa$, $|c| + 2 \leq \mu$, and $|a| + 1 \leq \mu$. With the aid of Lemma 7.3, we have achieved an upper bound of the form

$$C\langle t\rangle^{-3/2} E_\mu(u(t)) E^{1/2}_\kappa(u(t))$$

for the portion of the integrals over $r \leq c_0 t$ on the right of (8.3).

Away from the origin. It remains to estimate the right-hand side of (8.3) for $r \geq c_0 t$.

First, we consider the nonresonant terms, i.e., those for which $(i, j, k) \neq (k, k, k)$. If $i \neq j$ and $r \geq c_0 t$, then $\langle t\rangle^{3/2} \leq C\langle r\rangle \langle c_i t - r\rangle^{1/2} \langle c_j t - r\rangle$. Using (6.3) we have the estimate

$$\begin{aligned} \|\partial \Gamma^b u^i & \partial^2 \Gamma^c u^j \partial \Gamma^a u^k\|_{L^1(r \geq c_0 t)} \\ &\leq C\langle t\rangle^{-3/2} \|\langle r\rangle \langle c_i t - r\rangle^{1/2} \partial \Gamma^b u^i\|_{L^\infty} \|\langle c_j t - r\rangle \partial^2 \Gamma^c u^j\|_{L^2} \|\partial \Gamma^a u^k\|_{L^2} \\ &\leq C\langle t\rangle^{-3/2} \Big[E^{1/2}_{|b|+3}(u(t)) + \mathcal{X}_{|b|+3}(u(t))\Big] \mathcal{X}_{|c|+2}(u(t)) E^{1/2}_{|a|+1}(u(t)) \\ &\leq C\langle t\rangle^{-3/2} E_\mu(u(t)) E^{1/2}_\kappa(u(t)). \end{aligned}$$

Otherwise, if $j \neq k$, we pair the weight $\langle r\rangle \langle c_k t - r\rangle^{1/2}$ with $\partial \Gamma^a u^k$ in L^∞ to get the same upper bound.

We are left to consider the resonant terms in (8.3), i.e., $(i, j, k) = (k, k, k)$, in the region $r \geq c_0 t$. It is here, finally, where the null condition enters. An application of Lemma 5.1 yields the following upper bound for these terms:

$$\begin{aligned} C\langle t\rangle^{-1} \sum_k \sum_{\substack{b+c=a \\ |c| \leq \mu - 2}} \Big[&\|\Gamma^{b+1} u^k \partial^2 \Gamma^c u^k \partial \Gamma^a u^k\|_{L^1(r \geq c_0 t)} \\ &+ \|\partial \Gamma^b u^k \partial \Gamma^{c+1} u^k \partial \Gamma^a u^k\|_{L^1(r \geq c_0 t)} \\ &+ \|\langle c_k t - r\rangle \partial \Gamma^b u^k \partial^2 \Gamma^c u^k \partial \Gamma^a u^k\|_{L^1(r \geq c_0 t)} \Big]. \end{aligned}$$

We still need to squeeze out an additional decay factor of $\langle t\rangle^{-1/2}$.

Since $r \geq c_0 t$, we have $\langle r\rangle \geq C\langle t\rangle$. Thus, we have using (6.1) that

$$\begin{aligned}
&\|\Gamma^{b+1}u^k\partial^2\Gamma^c u^k\partial\Gamma^a u^k\|_{L^1(r\geq c_0t)}\\
&\quad\leq C\langle t\rangle^{-1/2}\|\langle r\rangle^{1/2}\Gamma^{b+1}u^k\|_{L^\infty(r\geq c_0t)}\|\partial^2\Gamma^c u^k\|_{L^2}\|\partial\Gamma^a u^k\|_{L^2}\\
&\quad\leq C\langle t\rangle^{-1/2}E^{1/2}_{|b|+3}(u(t))E_\mu(u(t))\\
&\quad\leq C\langle t\rangle^{-1/2}E^{1/2}_\kappa(u(t))E_\mu(u(t)).
\end{aligned}$$

In a similar fashion, the second term is handled using (6.2):

$$\begin{aligned}
&\|\partial\Gamma^b u^k\partial\Gamma^{c+1}u^k\partial\Gamma^a u^k\|_{L^1(r\geq c_0t)}\\
&\quad\leq C\langle t\rangle^{-1}\|\partial\Gamma^b u^k\|_{L^2}\|\langle r\rangle\partial\Gamma^{c+1}u^k\|_{L^\infty(r\geq c_0t)}\|\partial\Gamma^a u^k\|_{L^2}\\
&\quad\leq C\langle t\rangle^{-1}E^{1/2}_{|c|+3}(u(t))E_\mu(u(t))\\
&\quad\leq C\langle t\rangle^{-1}E^{1/2}_\kappa(u(t))E_\mu(u(t)).
\end{aligned}$$

The final set of terms are estimated using (6.2) again and (7.4a):

$$\begin{aligned}
&\|\langle c_kt{-}r\rangle\partial\Gamma^b u^k\partial^2\Gamma^c u^k\partial\Gamma^a u^k\|_{L^1(r\geq c_0t)}\\
&\quad\leq C\langle t\rangle^{-1}\|\langle r\rangle\partial\Gamma^b u^k\|_{L^\infty(r\geq c_0t)}\|\langle c_kt-r\rangle\partial^2\Gamma^c u^k\|_{L^2}\|\partial\Gamma^a u^k\|_{L^2}\\
&\quad\leq C\langle t\rangle^{-1}E^{1/2}_{|b|+3}(u(t))\mathcal{X}_{|c|+2}(u(t))E^{1/2}_\mu(u(t))\\
&\quad\leq C\langle t\rangle^{-1}E^{1/2}_\kappa(u(t))E_\mu(u(t)).
\end{aligned}$$

Combining all the estimates in this subsection, we obtain, thanks to (8.2), the following inequality for the lower energy:

$$\begin{aligned}
\widetilde{E}'_\mu(u(t)) &\leq C\langle t\rangle^{-3/2}E_\mu(u(t))E^{1/2}_\kappa(u(t))\\
&\leq C\langle t\rangle^{-3/2}\widetilde{E}_\mu(u(t))\widetilde{E}^{1/2}_\kappa(u(t)).
\end{aligned}\tag{8.6}$$

Conclusion of the proof. By (8.2), we have that the modified energy satisfies $\widetilde{E}^{1/2}_\mu(u(t)) \leq C\varepsilon$ for $0 \leq t < T_0$. So from (8.5), we find that

$$\widetilde{E}_\kappa(u(t)) \leq \widetilde{E}_\kappa(u(0))\langle t\rangle^{C\varepsilon},$$

provided ε is small. Inserting this bound into (8.6) and using (8.2), we obtain

$$\begin{aligned}
(1/2)E_\mu(u(t)) &\leq \widetilde{E}_\mu(u(t)) \leq \widetilde{E}_\mu(u(0))\exp CI\widetilde{E}^{1/2}_\kappa(u(0))\\
&\leq 2E_\mu(u(0))\exp 2CIE^{1/2}_\kappa(u(0)) < 2\varepsilon^2
\end{aligned}$$

with $I = \int_0^\infty\langle s\rangle^{-3/2+C\varepsilon}ds$. With this we see that $E^{1/2}_\mu(u(t))$ remains strictly less than 2ε throughout the closed interval $0 \leq t \leq T_0$. This shows that $E_\mu(u(t))$ is bounded for all time, which completes the proof of Theorem 3.1.

REFERENCES

[1] S. Alinhac, *The Null Condition for Quasilinear Wave Equations in Two Space Dimensions* I, preprint.

[2] S. Alinhac, *The Null Condition for Quasilinear Wave Equations in Two Space Dimensions* II, preprint.

[3] D. Christodoulou, *Global solutions of nonlinear hyperbolic equations for small initial data,* Comm. Pure Appl. Math., 39 (1986), pp. 267–282.

[4] 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.

[5] A. Hoshiga, *The lifespan of solutions to quasilinear hyperbolic systems in the critical case,* Funkcial. Ekvac., 41 (1998), pp. 167–188.

[6] A. Hoshiga and H. Kubo, *Global small amplitude solutions of nonlinear hyperbolic systems with a critical exponent under the null condition,* SIAM J. Math. Anal., 31 (2000), pp. 486–513.

[7] F. John, *Blow-up for quasilinear wave equations in three space dimensions,* Comm. Pure Appl. Math. 34 (1981), pp. 29–51.

[8] F. John and S. Klainerman, *Almost global existence to nonlinear wave equations in three space dimensions,* Comm. Pure Appl. Math., 37 (1984), pp. 443–455.

[9] 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.

[10] S. Katayama, *Global existence for systems of nonlinear wave equations in two space dimensions,* Publ. Res. Inst. Math. Sci., 29 (1993), pp. 1021–1041.

[11] S. Klainerman and G. Ponce, *Global, small amplitude solutions to nonlinear evolution equations,* Comm. Pure Appl. Math., 36 (1983), pp. 133–141.

[12] S. Klainerman, *Uniform decay estimates and the Lorentz invariance of the classical wave equation,* Comm. Pure Appl. Math., 38 (1985), pp. 321–332.

[13] S. Klainerman, *The null condition and global existence to nonlinear wave equations,* in Nonlinear Systems of Partial Differential Equations in Applied Mathematics, Part 1, Lectures in Appl. Math. 23, AMS, Providence, RI, 1986, pp. 293–326.

[14] S. Klainerman, *Remarks on the global Sobolev inequalities in the Minkowski space* $\mathbb{R}^{n+1}$, Comm. Pure Appl. Math., 40 (1987), pp. 111–117.

[15] S. Klainerman and T. Sideris, *On almost global existence for nonrelativistic wave equations in* $3D$, Comm. Pure Appl. Math., 49 (1996), pp. 307–321.

[16] M. Kovalyov, *Resonance-type behaviour in a system of nonlinear wave equations,* J. Differential Equations, 77 (1989), pp. 73–83.

[17] J. Shatah, *Global existence of small solutions to nonlinear evolution equations,* J. Differential Equations, 46 (1982), pp. 409–425.

[18] T. Sideris, *Global behavior of solutions to nonlinear wave equations in three dimensions,* Comm. Partial Differential Equations, 8 (1983), pp. 1291–1323.

[19] T. Sideris, *The null condition and global existence of nonlinear elastic waves,* Invent. Math., 123 (1996), pp. 323–342.

[20] T. Sideris, *Nonresonance and global existence of prestressed nonlinear elastic waves*, Ann. of Math. (2), 151 (2000), pp. 849–874.

[21] K. Yokoyama, *Global existence of classical solutions to systems of wave equations with critical nonlinearity in three space dimensions,* J. Math. Soc. Japan, 52 (2000), pp. 609–632.

SIAM J. MATH. ANAL.
Vol. 33, No. 2, pp. 489–507

INVARIANT MANIFOLDS AND LONG-TIME ASYMPTOTICS FOR THE VLASOV–POISSON–FOKKER–PLANCK EQUATION*

YOSHIYUKI KAGEI†

Dedicated to Professors Takaaki Nishida and Masayasu Mimura on their 60*th birthdays.*

Abstract. We study the large time behavior of small solutions to the Cauchy problem for the Vlasov–Poisson–Fokker–Planck equation, which is a degenerate parabolic equation with nonlocal nonlinearity. We construct finite dimensional invariant manifolds in a neighborhood of the origin in polynomially weighted Sobolev spaces, which enables us to compute systematically the long-time asymptotics for small solutions. To construct invariant manifolds, we make use of the "similarity variables" transformation as in C. E. Wayne's work in 1997, where invariant manifolds for parabolic equations in unbounded domains are constructed.

Key words. Vlasov–Poisson–Fokker–Planck equation, long-time asymptotics, invariant manifold

AMS subject classifications. 35B40, 35M99, 35Q99

PII. S0036141000371368

1. Introduction.

Many nonlinear parabolic partial differential equations often exhibit some scale-invariant structures in large time behavior of their solutions; namely, their solutions asymptotically have some self-similar profiles in large times. Based on this viewpoint, Bricmont, Kupiainen, and Lin [4] developed the renormalization group method to study large time behavior of solutions of certain types of nonlinear parabolic equations and derived the long-time asymptotics of solutions up to the leading order.

On the other hand, for many dissipative systems, the large time behavior of solutions is recognized to be controlled by a finite number of degrees of freedom. To study this point, invariant manifold theory is a strong mathematical tool, and, actually, the large time behavior of solutions is described by a system of a finite number of ordinary differential equations when the dimension of the constructed invariant manifold is finite. However, the application of the theory had been restricted to systems of ordinary differential equations and some classes of partial differential equations on bounded domains, where the spectra of linearized problems are discrete. In 1997, Wayne [21] constructed invariant manifolds for the problem

$$\partial_t f - \Delta_x f + F(f) = 0, \quad f = f(x,t), \quad x \in \boldsymbol{R}^N, \quad t > 0, \tag{1.1}$$

where $F(f) = O(|f|^p)$ as $|f| \to 0$ for some $p > 1$. To construct the invariant manifold for (1.1), Wayne used the change of variables to the so-called "similarity variables" and showed the existence of finite dimensional invariant manifolds in some Sobolev spaces with polynomial weights. As a result, for suitable F, asymptotic profiles of solutions in large time were given, up to orders higher than that given in [4]. Wayne's method has been generalized in a few directions: to the study of the long-time behavior of solutions around spatially periodic steady solutions of the Swift–Hohenberg equation

*Received by the editors April 20, 2000; accepted for publication (in revised form) March 14, 2001; published electronically July 31, 2001.

http://www.siam.org/journals/sima/33-2/37136.html

†Graduate School of Mathematics, Kyushu University, Fukuoka 812-8581, Japan (kagei@math.kyushu-u.ac.jp).

[11]; to problems on cylindrical domains [22]; to higher-order dissipative systems [10]. See also [15] for an approach in the L^p-framework. All of these problems have linearized parts of essentially diffusive type or of the type $(-\Delta_x)^n$.

The purpose of this paper is to carry out the same kind of analysis initiated by Wayne [21] for a different type of problem to study the long-time asymptotics of solutions to such problems. The problem discussed in this paper is the Cauchy problem for the Vlasov–Poisson–Fokker–Planck equation (without friction term)

$$\begin{aligned} &\partial_t f + u\cdot\nabla_x f + E(f)\cdot\nabla_u f - \Delta_u f = 0, \quad (x,u)\in \boldsymbol{R}^N\times\boldsymbol{R}^N,\ t>0,\\ &f|_{t=0} = f_0. \end{aligned} \tag{1.2}$$

Here $N\geq 2$, $f=f(x,u,t)$ is the unknown function, which describes the density of particles with respect to position $x\in\boldsymbol{R}^N$ and velocity $u\in\boldsymbol{R}^N$ at time t; $\nabla_x = (\partial_{x_1},\dots,\partial_{x_N})$, $\nabla_u=(\partial_{u_1},\dots,\partial_{u_N})$; Δ_u is the Laplacian with respect to the variable u: $\Delta_u = \partial_{u_1}^2+\cdots\partial_{u_N}^2$. $E(f)$ is an integral operator defined by

$$E(f)=\omega\frac{x}{|x|^N} *_x \int_{R^N} f(x,u,t)\,du, \quad \omega = +\frac{1}{\sigma^2|S^{N-1}|} \text{ or } -\frac{1}{\sigma^2|S^{N-1}|},$$

where $|S^{N-1}|$ is the $(N-1)$ dimensional volume of the N dimensional unit sphere; σ is a positive constant (called a diffusion constant); $*_x$ denotes the convolution with respect to x. Thus the equation in (1.2) is a degenerate parabolic equation with nonlocal nonlinearity.

We will construct finite dimensional invariant manifolds for (1.2) in some Sobolev spaces with polynomial weights and give long-time asymptotics of small solutions to (1.2). To state our main result we introduce function spaces

$$\begin{aligned} X_r^{\ell,m} = \{f(x,u)\in L^2(\boldsymbol{R}^N\times\boldsymbol{R}^N) : (1+|x|^2+|u|^2)^{r/2}\partial_x^\alpha\partial_u^\beta f \in L^2(\boldsymbol{R}^N\times\boldsymbol{R}^N),\\ 0\leq|\alpha|\leq\ell,\ 0\leq|\beta|\leq m\}, \end{aligned}$$

where ℓ, m, and r are nonnegative integers. The norm of $X_r^{\ell,m}$ is defined by

$$\|f\|_{X_r^{\ell,m}} = \left(\int\sum_{|\alpha|\leq\ell,|\beta|\leq m}(1+|x|^2+|u|^2)^r|\partial_x^\alpha\partial_u^\beta f(x,u)|^2\,dxdu\right)^{1/2}.$$

We now give the asymptotics of small solutions up to the order n.

THEOREM 1.1. *Let n be an integer satisfying $0\leq n\leq 3N-5$, and let r be an integer satisfying $r\geq n+3N+\frac{1}{2}$. Also, let m be an integer satisfying $m>N$. Then for any $\varepsilon>0$, if $\|f_0\|_{X_r^{m,m}}$ is sufficiently small, there exists a unique global solution $f(t)$ of (1.2) in $C([0,\infty);X_r^{m,m})$, and $f(t)$ satisfies*

$$\lim_{t\to\infty} t^{\frac{n+1}{2}-\varepsilon}\Big\|t^{2N}f(t^{3/2}x,t^{1/2}u,t)-\sum_{k=0}^{n}t^{-\frac{k}{2}}\sum_{3|\alpha|+|\beta|=k}B_{\alpha,\beta}g_{\alpha,\beta}(x,u)\Big\|_{L^\infty_{x,u}}=0.$$

Here $g_{\alpha,\beta}(x,u)=c_{\alpha,\beta}\partial_x^\alpha(\partial_x+\partial_u)^\beta e^{-3|x-\frac{u}{2}|^2-\frac{1}{4}|u|^2}$ and $c_{\alpha,\beta}=(-\frac{1}{3})^{|\alpha|}\frac{1}{\alpha!\beta!}(\frac{\sqrt{3}}{2\pi})^N$ with α and β being multi-indices; and $B_{\alpha,\beta}$ are constants determined by f_0 and the nonlinearity. In particular, $B_{0,0}=\int f_0(x,u)\,dxdu$.

Remark 1.2. In Theorem 1.1 the range of n is restricted as $0 \leq n \leq 3N-5$. One can, however, obtain the asymptotics of f up to any nonnegative $n \in \mathbf{Z}$; in fact, the full dynamics have been reduced to those of the system of ordinary differential equations (2.6) below.

If n is beyond the range in Theorem 1.1, i.e., if $n \geq 3N-4$, then the effect of the nonlinearity becomes somewhat stronger, and logarithmic terms appear in the asymptotics. For example, if $n = 3N-4$, then we have

$$\begin{aligned}
&t^{2N} f(t^{3/2}x, t^{1/2}u, t) \\
&\quad \sim \sum_{k=0}^{n-1} \sum_{3|\alpha|+|\beta|=k} (B_{\alpha,\beta} t^{-\frac{k}{2}} + \widetilde{B}_{\alpha,\beta} t^{-\frac{n}{2}}) g_{\alpha,\beta}(x,u) \\
&\qquad + \sum_{3|\alpha|+|\beta|=n} (B_{\alpha,\beta} t^{-\frac{n}{2}} + \widetilde{B}_{\alpha,\beta} t^{-\frac{n}{2}} \log t)\, g_{\alpha,\beta}(x,u) \\
&\qquad + h(x,u,t) + O(t^{-\frac{n+1}{2}+\varepsilon}), \quad (n = 3N-4),
\end{aligned}$$

where $B_{\alpha,\beta}$ and $\widetilde{B}_{\alpha,\beta}$ are some constants and $h(x,u,t) = O(t^{-\frac{n}{2}})$. (For example, when $N = 2$, the constants $\widetilde{B}_{0,\beta}$ ($|\beta| = 2$) are given by $\widetilde{B}_{0,\beta} = -\frac{3}{8} {B_{0,0}}^2 \omega$ for $\beta = (2,0)$ and $(0,2)$, $\widetilde{B}_{0,\beta} = 0$ for $\beta = (1,1)$.) As for the function $h(x,u,t)$, see section 2.

The existence, uniqueness, and regularity of solutions to (1.2) have been widely studied; see, e.g., [1, 2, 3, 5, 7, 8, 9, 17, 18, 19, 20] and references therein. Among these works, this paper is closely related to those in [5, 9, 17]. The work [9] by Carrillo, Soler, and Vázquez is the first one among the works for (1.2) to make use of the scaling-invariant property of the fundamental solution of the linearized problem. It is shown in [9] that under some conditions the long-time asymptotics of weak solutions can be obtained up to the leading order when $N \geq 3$. An existence of solutions satisfying these conditions was shown in [8], and at least for initial data small enough in some sense, such solutions exist. Carpio [5] then studied large time behavior of small solutions when $N = 3$. In the analysis in [5] the fundamental solution of the linearized problem of (1.2) was investigated in detail and the long-time asymptotics was obtained up to the second order by using the rescaling technique. The result in [5] shows that the effect of the nonlinearity appears in the second order term of the asymptotics in a weak sense. Ono and Strauss [17] recently obtained sharp decay estimates for the difference of the solution to (1.2) and the solution of the corresponding linearized problem for any $N \geq 2$ if the initial value is sufficiently small. Our results extend these results to higher-order asymptotics and indicate the order in the asymptotics where the effect of the nonlinearity becomes strong.

We prove Theorem 1.1 and Remark 1.2 by constructing finite dimensional invariant manifolds as in [10, 11, 21, 22]. We change the variables into the "similarity" variables:

$$\begin{aligned}
&\tilde{t} = \log(t+1), \quad \tilde{x} = x/(t+1)^{3/2}, \quad \tilde{u} = u/(t+1)^{1/2}, \\
&f(x,u,t) = (t+1)^{-2N} \tilde{f}(x/(t+1)^{3/2}, u/(t+1)^{1/2}, \log(t+1))
\end{aligned}$$

(cf. [9]). Then the equation for $\tilde{f}$ is written, after omitting tildes, as

(1.3)
$$\begin{aligned}
&\partial_t f - (\tfrac{3}{2}x - u) \cdot \nabla_x f - \tfrac{1}{2} u \cdot \nabla_u f - 2Nf + e^{-(\frac{3}{2}N-2)t} E(f) \cdot \nabla_u f - \Delta_u f = 0, \\
&f|_{t=0} = f_0.
\end{aligned}$$

In [10, 11, 21] (see also [15]), to construct invariant manifolds in polynomially weighted spaces, the crucial step is, roughly speaking, to show that the linearized semigroup $T(t)$ in the similarity variables behaves as

$$\begin{cases} P_jT(t)f_0 = e^{\lambda_k t}P_jf_0, \quad k = 0, \dots, n, \\ \|Q_nT(t)f_0\| \le Ce^{\lambda_{n+1}t}\|f_0\| \end{cases} \tag{1.4}$$

in some spaces with polynomial weights. Here λ_k, $k = 0, 1, \dots, n$ are the first $n + 1$ eigenvalues of the linearized operator with $\lambda_n < \cdots < \lambda_0$; P_k denotes the eigenprojection associated with the eigenvalue λ_k; $Q_n = I - \sum_{k=0}^{n} P_k$; and λ_{n+1} is a number satisfying $\lambda_{n+1} < \lambda_n$. The proof of (1.4) in [10, 11, 21] is done by a skillful decomposition of the underlying domain $\boldsymbol{R}^N$ and a use of the fact that the growth bound for $Q_nT(t)$ in an exponentially weighted space can be estimated by the spectral bound of its generator. It is the same in [15]. The method in [10, 11, 21] also works in our case since the spectrum of our $T(t)$ in some exponentially weighted space consists only of discrete eigenvalues. (See the remarks in the appendix.) However, in this paper we will show (1.4) by directly analyzing the Fourier transform of our $T(t)$ in polynomially weighted spaces. The "similarity-variable" transformation plays an important role in bringing out clearly the discrete nature of the spectrum of $T(t)$, which can be seen more easily through the Fourier transform of $T(t)$. We derive a useful expression of a spectral representation of $T(t)f_0$ in terms of its Fourier transform. This expression naturally leads us to the behavior of the semigroup as in (1.4) without analysis of the linearized problem in the exponentially weighted space. The dissipative nature of the problem works well in controlling various quantitative estimates. The method is also applicable for the case (1.1). It is still unclear to the author what kind of structures of asymptotic self-similarity and dissipativity are needed for the analysis by the invariant manifold method on unbounded domains as initiated by Wayne [21].

The paper is organized as follows. In section 2 we reformulate the problem in the similarity variables. The existence theorem (Theorem 2.1) of invariant manifolds for (1.3) is then stated, and the proofs of Theorem 1.1 and Remark 1.2 are outlined. Some comments on the case $N = 1$ are given in Remark 2.2. In section 3 we investigate some spectral properties of the linearized operator and show the behavior of the linearized semigroup as in (1.4). In section 4 we derive some estimates for the nonlinearity, which, together with (1.4), implies Theorem 2.1. The appendix is devoted to a derivation of an integral formula of the linearized semigroup for (1.3).

2. Formulation of the problem in the similarity variables. In this section we reformulate the problem in the similarity variables. We then present the existence theorem of the invariant manifolds (Theorem 2.1), and the proofs of Theorem 1.1 and Remark 1.2 are outlined.

Let us transform the problem into the one in the similarity variables. We change the variables as

$$\begin{aligned} &\tilde{t} = \log(t+1), \quad \tilde{x} = x/(t+1)^{3/2}, \quad \tilde{u} = u/(t+1)^{1/2}, \\ &f(x,u,t) = (t+1)^{-\gamma}\tilde{f}(x/(t+1)^{3/2}, u/(t+1)^{1/2}, \log(t+1)), \end{aligned}$$

where $\gamma = \frac{1}{2}N + 2$. The change of variables here is slightly different from the ones written in the introduction. Since the nonlinearity in (1.2) is just quadratic, this transformation is convenient, and the transformed problem becomes autonomous. In fact, the nonlinearity is transformed as

$$E(f)\cdot\nabla_u f(x,u,t) = (t+1)^{-2\gamma+\frac{N}{2}+1}E(\tilde{f})\cdot\nabla_{\tilde{u}}\tilde{f}(x/(t+1)^{3/2}, u/(t+1)^{1/2}, \log(t+1)),$$

and the equation for $\tilde{f}$ is written, after omitting tildes, as

$$\begin{aligned}(2.1)\qquad &\partial_t f - \left(\frac{3}{2}x - u\right)\cdot\nabla_x f - \frac{1}{2}u\cdot\nabla_u f - \gamma f + E(f)\cdot\nabla_u f - \Delta_u f = 0,\\ &f|_{t=0} = f_0.\end{aligned}$$

We write the problem (2.1) in the form

$$\partial_t f = Lf - \mathcal{N}(f), \quad f(0) = f_0,$$

where $Lf = \Delta_u f + (\frac{3}{2}x - u)\cdot\nabla_x f + \frac{1}{2}u\cdot\nabla_u f + \gamma f$ and $\mathcal{N}(f) = E(f)\cdot\nabla_u f$.

We first consider the linear problem in the weighted space $X_r^{\ell,m}$. As we will see in section 3, the linearized operator has the following properties.

We are given a nonnegative integer n, and we fix this n hereafter. For this n we take the weight large enough in such a way that $r \geq n + 3N + \frac{1}{2}$. Then, as for the spectrum $\sigma(L)$ of L in $X_r^{0,0}$, we have

$$\sigma(L) \subset \{\lambda_k : k = 0, 1, \dots, n\} \cup \{\mathrm{Re}\lambda \leq \lambda_{n+1}\} \quad \left(\lambda_j = -(2N-\gamma) - \frac{j}{2}\right).$$

Here each of the λ_k $(k = 0, 1, \dots, n)$ is a semisimple eigenvalue; the associated eigenspace is spanned by functions $g_{\alpha,\beta}$'s with α and β satisfying $3|\alpha| + |\beta| = k$, where

$$\begin{aligned}&g_{\alpha,\beta}(x,u) = c_{\alpha,\beta}\partial_x^\alpha(\partial_x + \partial_u)^\beta e^{-\mu(x,u)}, \quad \mu(x,u) = 3\left|x - \frac{u}{2}\right|^2 + \frac{1}{4}|u|^2,\\ &c_{\alpha,\beta} = \left(-\frac{1}{3}\right)^{|\alpha|}\frac{1}{\alpha!\beta!}\left(\frac{\sqrt{3}}{2\pi}\right)^N.\end{aligned}$$

The eigenprojection P_k associated with λ_k is given by

$$P_k f = \sum_{3|\alpha|+|\beta|=k}\langle f, g^*_{\alpha,\beta}\rangle g_{\alpha,\beta}.$$

Here $g^*_{\alpha,\beta}(x,u) = (\partial_x + 3\partial_u)^\alpha(\partial_x + 2\partial_u)^\beta e^{-\mu(x,u)}$ denotes the adjoint eigenfunction, and the pairing $\langle\cdot,\cdot\rangle$ is defined by

$$\langle f, g\rangle = \int f(x,u)g(x,u)e^{\mu(x,u)}\,dxdu.$$

Note that $\langle g_{\alpha,\beta}, g^*_{\tilde{\alpha},\tilde{\beta}}\rangle = 1$ if $(\alpha,\beta) = (\tilde{\alpha},\tilde{\beta})$, and $\langle g_{\alpha,\beta}, g^*_{\tilde{\alpha},\tilde{\beta}}\rangle = 0$ if $(\alpha,\beta) \neq (\tilde{\alpha},\tilde{\beta})$. We denote by $\mathcal{P}_n = \sum_{k=0}^n P_k$ the projection onto the spectral subspace corresponding to discrete eigenvalues $\{\lambda_k\}_{k=0}^n$, and we define $\mathcal{Q}_n$ by $\mathcal{Q}_n = I - \mathcal{P}_n$.

By Proposition 3.3 below, $\mathcal{P}_n$ is a bounded operator in $X_r^{\ell,m}$ since $r \geq n + 3N + \frac{1}{2}$ and $X_r^{\ell,m}$ is decomposed into the direct sum

$$X_r^{\ell,m} = Y_n \oplus Z_{r,n}^{\ell,m}, \quad Y_n \equiv \mathcal{P}_n X_r^{\ell,m}, \quad Z_{r,n}^{\ell,m} \equiv \mathcal{Q}_n X_r^{\ell,m},$$

and the solution $T(t)f_0$ of the linear problem is decomposed as

$$\begin{aligned}&T(t)f_0 = \varphi_n(t) + \psi(t), \quad \varphi_n(t) \in Y_n, \quad \psi(t) \in Z_{r,n}^{\ell,m},\\ &\varphi_n(t) = \sum_{k=0}^n e^{\lambda_k t}P_k f_0, \quad \psi(t) = \mathcal{Q}_n T(t)f_0.\end{aligned}$$

As for the part $\psi(t) = \mathcal{Q}_n T(t) f_0$ on the subspace $Z^{\ell,m}_{r,n}$, the estimate

$$\|\mathcal{Q}_n T(t) f_0\|_{X^{\ell,m}_r} \leq C(1 + t^{-\frac{j}{2}}) e^{\lambda_{n+1} t} \|f_0\|_{X^{\ell,m-j}_r} \tag{2.2}$$

holds for $\ell \geq 0$, $m \geq j$, and $j = 0, 1$. Therefore, the large time behavior of solutions of the linear problem is described, up to $O(e^{\lambda_{n+1} t})$, by the behavior of solutions on the *finite dimensional invariant subspace* Y_n.

For the nonlinear problem we have the following theorem, from which the long-time asymptotics given in Theorem 1.1 and Remark 1.2 are obtained.

Theorem 2.1. *Let $n \geq 0$ be an integer, and let r be an integer satisfying $r \geq n + 3N + \frac{1}{2}$. Then for any fixed integers $m \geq 0$, $j \geq 1$, and $\ell = [\frac{N}{2} - 1] + 1$, there exists a finite dimensional invariant manifold $\mathcal{M}$ for* (2.1) *in a neighborhood of the origin of $X^{m+\ell,j}_r$, i.e., there exist $\Phi \in C^1(Y_n; Z^{m+\ell,j}_{r,n})$ and $R > 0$ such that $\Phi(0) = 0$, $D\Phi(0) = 0$, and*

$$\mathcal{M} = \{\varphi_n + \Phi(\varphi_n); \varphi_n \in Y_n, \|\varphi_n\|_{X^{m+\ell,j}_r} \leq R\},$$

where $Y_n = \mathcal{P}_n X^{m+\ell,j}_r$ and $Z^{m+\ell,j}_{r,n} = \mathcal{Q}_n X^{m+\ell,j}_r$; and $\mathcal{M}$ is invariant under semiflows defined by (2.1)*. Furthermore, solutions near the origin stay in a neighborhood of the origin for all times and approach to $\mathcal{M}$ at a rate $O(e^{(\lambda_{n+1}+\varepsilon)t})$ as $t \to \infty$. More precisely, if $\|f_0\|_{X^{m+\ell,j}_r}$ is sufficiently small, then there uniquely exists a solution $\bar{f}(t)$ of* (2.1) *on $\mathcal{M}$ such that*

$$\|f(t) - \bar{f}(t)\|_{X^{m+\ell,j}_r} \leq C e^{(\lambda_{n+1}+\varepsilon)t}. \tag{2.3}$$

Theorem 2.1 follows from Propositions 3.6 and 4.1 below by applying standard arguments of invariant manifold theory in [6, 16].

We now outline how to obtain the long-time asymptotics given in Theorem 1.1 and Remark 1.2.

Our starting point is the estimate (2.3) in Theorem 2.1. We can rewrite the estimate (2.3) in the form

$$\|\varphi_n(t) - \bar{\varphi}_n(t)\|_{X^{m+\ell,j}_r} \leq C e^{(\lambda_{n+1}+\varepsilon)t} \tag{2.4}$$

and

$$\|\psi(t) - \Phi(\bar{\varphi}_n(t))\|_{X^{m+\ell,j}_r} \leq C e^{(\lambda_{n+1}+\varepsilon)t}, \tag{2.5}$$

where

$$f(t) = \varphi_n(t) + \psi(t), \quad \bar{f}(t) = \bar{\varphi}_n(t) + \Phi(\bar{\varphi}_n(t)), \quad \varphi_n(t), \bar{\varphi}_n(t) \in Y_n, \quad \psi(t) \in Z^{m+\ell,j}_{r,n}.$$

From (2.4) and (2.5) we can see that to obtain the asymptotics of $f(t)$ up to $O(e^{(\lambda_{n+1}+\varepsilon)t})$, it suffices to investigate the behavior of $\bar{\varphi}_n(t)$, which is governed by a system of a finite number of ordinary differential equations. Since $\bar{\varphi}_n(t)$ can be written as

$$\bar{\varphi}_n(t) = \sum_{3|\alpha|+|\beta| \leq n} \varphi_{\alpha,\beta}(t) g_{\alpha,\beta}, \quad \varphi_{\alpha,\beta} \in \mathbf{R},$$

the problem is reduced to the analysis of the behavior of $\varphi_{\alpha,\beta}$'s.

We now derive a system of ordinary differential equations for $\varphi_{\alpha,\beta}$'s. Since $\bar{f}(t) = \bar{\varphi}_n(t) + \Phi(\bar{\varphi}_n(t))$ is a solution of (2.1) on $\mathcal{M}$, it satisfies

$$\partial_t \bar{f} = L\bar{f} - \mathcal{N}(\bar{f}), \quad \bar{f} = \bar{\varphi}_n + \Phi(\bar{\varphi}_n).$$

Taking the pairing $\langle\cdot,\cdot\rangle$ of this identity with $g^*_{\alpha,\beta}$, we have

$$\dot{\varphi}_{\alpha,\beta} = \lambda_k \varphi_{\alpha,\beta} + H_{\alpha,\beta}(\bar{\varphi}_n), \quad 3|\alpha| + |\beta| = k,\ 0 \le k \le n, \tag{2.6}$$

where $\dot{\varphi}_{\alpha,\beta} = \frac{d}{dt}\varphi_{\alpha,\beta}$ and $H_{\alpha,\beta}(\bar{\varphi}_n) = -\langle \mathcal{N}(\bar{\varphi}_n + \Phi(\bar{\varphi}_n)), g^*_{\alpha,\beta}\rangle$.

For $\alpha = \beta = 0$, one can easily verify that $H_{0,0}(\bar{\varphi}_n) = 0$. Hence

$$\dot{\varphi}_{0,0} = \lambda_0 \varphi_{0,0}, \quad \text{i.e.,} \quad \varphi_{0,0}(t) = e^{\lambda_0 t}\varphi_{0,0}(0).$$

Recall that $\lambda_0 = -(2N - \gamma) = -(\frac{3}{2}N - 2) < 0$. For $(\alpha, \beta) \neq (0, 0)$, we have, by the variation of constants formula,

$$\varphi_{\alpha,\beta}(t) = e^{\lambda_k t}\varphi_{\alpha,\beta}(0) + e^{\lambda_k t}\int_0^t e^{-\lambda_k s} H_{\alpha,\beta}(\bar{\varphi}_n(s))\, ds \tag{2.7}$$

with $k = 3|\alpha| + |\beta|$, $1 \le k \le n$. Since $\lambda_k = \lambda_0 - \frac{k}{2}$, one can expect that $\varphi_{\alpha,\beta}(t)$ decays strictly faster than $\varphi_{0,0}(t)$. Therefore, the slowest term in $H_{\alpha,\beta}(\bar{\varphi}_n(s))$ behaves like $e^{2\lambda_0 s}$, since the lowest order terms of $H_{\alpha,\beta}(\bar{\varphi}_n)$ are quadratic in $\{\varphi_{\alpha,\beta}\}$. As a result, the integrand in (2.7) behaves like $e^{(2\lambda_0 - \lambda_k)s}$.

Now let $n \le 3N - 5$. This is just equivalent to $|\lambda_n| < 2|\lambda_0|$ (and to $|\lambda_{n+1}| \le 2|\lambda_0|$). It then follows that for $3|\alpha| + |\beta| = k,\ 0 \le k \le n$,

$$\varphi_{\alpha,\beta}(t) \sim \text{const. } e^{\lambda_k t} + O(e^{2\lambda_0 t}),$$

where const. depends on $\varphi_{\alpha,\beta}(0)$ and $H_{\alpha,\beta}$. We can also obtain

$$\|\psi(t)\|_{X_r^{m+\ell,j}} \le C e^{(\lambda_{n+1}+\varepsilon)t}.$$

Therefore,

$$\tilde{f}(x, u, \tilde{t}) \sim \sum_{k=0}^{n} e^{\lambda_k \tilde{t}} \sum_{3|\alpha|+|\beta|=k} B_{\alpha,\beta}\, g_{\alpha,\beta}(x, u) + O(e^{(\lambda_{n+1}+\varepsilon)\tilde{t}}). \tag{2.8}$$

Here we write the solution of (2.1) and the time variable with tildes. Now, converting (2.8) to the original function f and the original time scale t, we obtain the asymptotics given in Theorem 1.1 for $n \le 3N - 5$ if we choose m and j as $m + l > N$ and $j > N$. Note that this choice of m and j implies that $X_r^{m+\ell,j} \subset L^\infty(dxdu)$ due to the Sobolev embedding.

We next consider higher-order asymptotics. In higher-order cases, the estimates (2.4), (2.5), and (2.6) for $\varphi_{\alpha,\beta}$'s, of course, take the same forms. Let $n \ge 3N - 4$. Then $|\lambda_n| \ge 2|\lambda_0|$ and $|\lambda_{n+1}| > 2|\lambda_0|$. Therefore, the integrand in (2.7) does not decay as $s \to \infty$ for some α and β, and the effect of the inhomogeneous term is no longer weak. Also, one must take the effect of $\Phi(\bar{\varphi}_n(t))$ into account, and thus the form of the asymptotics becomes complicated.

For example, if $n = 3N - 4$, then we have $\lambda_n = 2\lambda_0$, and, therefore, the integrand in (2.7) with $3|\alpha| + |\beta| = n$ is of $O(1)$. It then follows that for $3|\alpha| + |\beta| = n$,

$$\varphi_{\alpha,\beta}(t) \sim c_1 e^{\lambda_n t} + c_2\, t e^{\lambda_n t} + O(e^{\lambda_{n+1} t}),$$

where c_1 and c_2 are some constants. One can also see that $\Phi(\bar{\varphi}_n(t)) = O(e^{\lambda_n t})$. Combining these with (2.4) and (2.5) and converting to the original function f and the original time scale t, we see that

$$\begin{aligned}
&t^{2N} f(t^{3/2}x, t^{1/2}u, t) \\
&\sim \sum_{k=0}^{n-1} \sum_{3|\alpha|+|\beta|=k} (B_{\alpha,\beta} t^{-\frac{k}{2}} + \widetilde{B}_{\alpha,\beta} t^{-\frac{n}{2}}) g_{\alpha,\beta}(x,u) \\
&\quad + \sum_{3|\alpha|+|\beta|=n} (B_{\alpha,\beta} t^{-\frac{n}{2}} + \widetilde{B}_{\alpha,\beta} t^{-\frac{n}{2}} \log t)\, g_{\alpha,\beta}(x,u) \\
&\quad + h(x,u,t) + O(t^{-\frac{n+1}{2}+\varepsilon}) \quad (n = 3N-4),
\end{aligned}$$

where $B_{\alpha,\beta}$ and $\widetilde{B}_{\alpha,\beta}$ are some constants and $h(x,u,t) = O(t^{-\frac{n}{2}})$. This gives the asymptotics presented in Remark 1.2 for $n = 3N - 4$. (A direct calculation, for example, for $N = 2$ shows that the constants $\widetilde{B}_{0,\beta}$ ($|\beta| = 2$) are given by $\widetilde{B}_{0,\beta} = -\frac{3}{8} {B_{0,0}}^2 \omega$ for $\beta = (2,0)$ and $(0,2)$ and $\widetilde{B}_{0,\beta} = 0$ for $\beta = (1,1)$.) For $n \geq 3N - 3$, it is possible to obtain the asymptotics in a similar manner as above, but the form of the asymptotics becomes more complicated.

Remark 2.2. In the case when $N = 1$, the nonlinearity is relevant [4, 10]. However, as in [10], if $\int f_0 \, dxdu = 0$, one may obtain something about the dynamics of small solutions. But in this case the dynamics will strongly depend on the sign of ω (in $E(f)$), and we do not consider this case here.

In the remaining part of this paper we will prove Theorem 2.1. The strategy of the proof is similar to that in [10, 11, 21], and the theorem follows from Propositions 3.6 and 4.1 below by applying standard arguments of invariant manifold theory in [6, 16].

3. Some spectral properties of the linearized operator. In this section we deduce some spectral properties of the linearized operator, which is needed for the construction of invariant manifolds. We first investigate the spectrum of the operator A:

$$Af = \left(\frac{3}{2}x - u\right) \cdot \nabla_x f + \frac{1}{2} u \cdot \nabla_u f + 2Nf + \Delta_u f.$$

Note that the linearized operator L which appears in (2.1) is given by $L = A - (2N - \gamma)I$.

Let n be a nonnegative integer. We fix this n hereafter and take an integer r satisfying $r \geq n + 3N + \frac{1}{2}$. We discuss the spectrum of A in $X_r^{0,0}$. The Fourier transform of Af is given by

$$\hat{A}\hat{f} \equiv \widehat{Af} = -\frac{3}{2}\xi \cdot \nabla_\xi \hat{f} - \left(\frac{1}{2}w - \xi\right) \cdot \nabla_w \hat{f} - |w|^2 \hat{f},$$

where $\hat{f}(\xi, w)$ is the Fourier transform of $f(x,u)$:

$$\hat{f}(\xi, w) = \int e^{-ix\cdot\xi - iu\cdot w} f(x,u)\, dxdu.$$

Consider now the eigenvalue problem $\lambda \hat{f} - \hat{A}\hat{f} = 0$. Transforming the variables

$$\begin{aligned}
&\tilde{w} = w + \xi, \quad \tilde{\xi} = \xi, \\
&\hat{f}(\xi, w) = \hat{g}(\xi, w + \xi)\, e^{-\hat{\mu}(\xi,w)}, \quad \hat{\mu}(\xi, w) = |w + \tfrac{\xi}{2}|^2 + \tfrac{1}{12}|\xi|^2,
\end{aligned}$$

we reduce the problem to

$$\lambda\hat{g} + \frac{3}{2}\tilde{\xi}\cdot\nabla_{\tilde{\xi}}\hat{g} + \frac{1}{2}\tilde{w}\cdot\nabla_{\tilde{w}}\hat{g} = 0.$$

On can easily verify that each $\sigma_k = -\frac{k}{2}$ $(k = 0, 1, 2\cdots)$ is an eigenvalue, and the associated eigenfunctions are given by $\tilde{\xi}^\alpha \tilde{w}^\beta$ with multi-indices α and β satisfying $3|\alpha| + |\beta| = k$, and in the original variables, $\xi^\alpha(w+\xi)^\beta e^{-\hat{\mu}(\xi,w)}$. Therefore, taking the inverse Fourier transform, we have the following.

Proposition 3.1. *The spectrum $\sigma(A)$ of A in $X_r^{0,0}$ contains a set of eigenvalues $\{\sigma_k = -\frac{k}{2} : k = 0, 1, 2, \cdots\}$, and eigenfunctions associated with σ_k are given by $g_{\alpha,\beta}$ with multi-indices α and β satisfying $3|\alpha| + |\beta| = k$.*

We next consider the adjoint problem with respect to the pairing $\langle\cdot,\cdot\rangle$. The adjoint operator A^* of A with respect to the pairing $\langle\cdot,\cdot\rangle$ is given by

$$A^* = -\left(\frac{3}{2}x - u\right)\cdot\nabla_x + \left(\frac{7}{2}u - 6x\right)\cdot\nabla_u + 2N + \Delta_u.$$

As above, we have the following proposition.

Proposition 3.2. *The spectrum $\sigma(A^*)$ of A^* in $X_r^{0,0}$ contains a set of eigenvalues $\{\sigma_k = -\frac{k}{2} : k = 0, 1, 2, \cdots\}$, and eigenfunctions associated with σ_k are given by $g^*_{\alpha,\beta}$ with multi-indices α and β satisfying $3|\alpha| + |\beta| = k$.*

Since the linearized operator L of (2.1) is written as $L = A - (2N-\gamma)I$, we see that $\lambda_k = \sigma_k - (2N-\gamma)$ $(k = 0, 1, 2, \cdots)$ are eigenvalues of L. The projection $\mathcal{P}_n$ onto the eigenspaces of the first $n+1$ eigenvalues of L is formally given by

$$\mathcal{P}_n f = \sum_{k=0}^{n} P_k f = \sum_{k=0}^{n}\sum_{3|\alpha|+|\beta|=k}\langle f, g^*_{\alpha,\beta}\rangle g_{\alpha,\beta}.$$

As in [21], the following proposition shows that $\mathcal{P}_n$ is well defined in $X_r^{0,0}$ if r is suitably large.

Proposition 3.3. *If $r \geq n + \frac{1}{2}(2N+1)$, then $\mathcal{P}_n$ is well defined as a bounded operator in $X_r^{0,0}$.*

Proof. Let α and β satisfy $3|\alpha| + |\beta| \leq n$. Since $|g^*_{\alpha,\beta}(x,u)| \leq C(1+|x|+|u|)^n e^{-\mu(x,u)}$, we have

$$\begin{aligned}|\langle f, g^*_{\alpha,\beta}\rangle| &\leq C\int(1+|x|+|u|)^n|f|\,dxdu\\ &= C\int(1+|x|+|u|)^{-(2N+1)/2}(1+|x|+|u|)^{n+(2N+1)/2}|f|\,dxdu\\ &\leq C\left(\int(1+|x|+|u|)^{2n+(2N+1)}|f|^2\,dxdu\right)^{1/2},\end{aligned}$$

and the proposition follows. □

We next give a useful representation of the Fourier transform of $P_k f$.

Lemma 3.4. *Let k be a nonnegative integer, and let r be an integer satisfying $r \geq k + \frac{1}{2}(2N+1)$. Then*

$$\widehat{P_k f}(\xi,w) = \sum_{3|\alpha|+|\beta|=k}(3i)^{|\alpha|}(-i)^{|\beta|}[(\partial_\xi - \partial_w)^\alpha\partial_w^\beta(\hat{f}e^{\hat{\mu}})]\Big|_{\xi=w=0}\hat{g}_{\alpha,\beta}(\xi,w),$$

where $\hat{\mu} = \hat{\mu}(\xi,w) = |w+\frac{\xi}{2}|^2 + \frac{1}{12}|\xi|^2$.

The lemma can be shown by a simple application of the induction argument on α and β. So we omit the proof.

We next discuss the semigroup generated by A. As we will see in the appendix, the Fourier transform of the semigroup $S(t)$ generated by A takes the form

$$\widehat{S(t)f_0}(\xi,w)=\hat{f}(\xi,w,t)=\hat{f}_0(e^{-\frac{3}{2}t}\xi,e^{-\frac{1}{2}t}w+e^{-\frac{1}{2}t}a(t)\xi)e^{-a(t)|w+\frac{a(t)}{2}\xi|^2-\frac{a(t)^3}{12}|\xi|^2}.$$

Taking the inverse Fourier transform then gives an integral formula of the semigroup $S(t)f_0$:

$$S(t)f_0=\left(\tfrac{\sqrt{3}}{2\pi}\right)^N a(t)^{-2N}\int e^{-\frac{3|x-e^{-\frac{3}{2}t}y-\frac{a(t)}{2}(u+e^{-\frac{t}{2}}v)|^2}{a(t)^3}-\frac{|u-e^{-\frac{t}{2}}v|^2}{4a(t)}}f_0(y,v)dydv,$$

where $a(t)=1-e^{-t}$. See the appendix for the derivation of the formula of $S(t)f_0$.

Proposition 3.5. *Let r be an integer satisfying $r\geq n+\frac{1}{2}(2N+1)$. Then*

$$(\widehat{P_kS(t)}f)(\xi,w)=e^{-\hat{\mu}(\xi,w)}\sum_{3|\alpha|+|\beta|=k}\tfrac{1}{\alpha!\beta!}\partial_{\tilde{\xi}}^{\alpha}\partial_{\tilde{w}}^{\beta}F(0,0)(e^{-\frac{3}{2}t}\xi)^{\alpha}(e^{-\frac{t}{2}}(w+\xi))^{\beta}.$$

Here $F(\tilde{\xi},\tilde{w})=e^{\tilde{\mu}(\tilde{\xi},\tilde{w})}\hat{f}(\tilde{\xi},\tilde{w}-\tilde{\xi})$ and $\tilde{\mu}(\tilde{\xi},\tilde{w})=|\tilde{w}-\frac{\tilde{\xi}}{2}|^2+\frac{1}{12}|\tilde{\xi}|^2$.

Proof. Since $a(t)=1-e^{-t}$, we have

$$a(t)\left|w+\frac{a(t)}{2}\xi\right|^2+\frac{a(t)^3}{12}|\xi|^2=\hat{\mu}(\xi,w)-\tilde{\mu}(e^{-\frac{3}{2}t}\xi,e^{-\frac{1}{2}t}(w+\xi)),$$

and so

$$\widehat{S(t)f}=\hat{f}(e^{-\frac{3}{2}t}\xi,e^{-\frac{1}{2}t}(w+\xi)-e^{-\frac{3}{2}t}\xi)e^{\tilde{\mu}(e^{-\frac{3}{2}t}\xi,e^{-\frac{1}{2}t}(w+\xi))-\hat{\mu}(\xi,w)}.$$

Lemma 3.4 then applies to yield the desired formula since $\tilde{\mu}(\tilde{\xi},\tilde{w})=\hat{\mu}(\xi,w)$ and $\partial_{\tilde{w}}=\partial_w$, $\partial_{\tilde{\xi}}=\partial_\xi-\partial_w$ under the transformation $\tilde{\xi}=\xi$ and $\tilde{w}=w+\xi$. This completes the proof. □

Since the semigroup $T(t)$ generated by $L=A-(2N-\gamma)I$ is written as $T(t)=e^{-(2N-\gamma)t}S(t)$, Proposition 3.6 shows that $T(t)$ behaves as in (1.4) in $X_r^{0,0}$ if $r\geq n+3N+\frac{1}{2}$. In particular, for the spectrum $\sigma(L)$ of L in $X_r^{0,0}$, it holds that

$$\sigma(L)\subset\{\lambda_k\,:\,k=0,1,\dots,n\}\cup\{\operatorname{Re}\lambda\leq\lambda_{n+1}\}$$

if $r\geq n+3N+\frac{1}{2}$; each of λ_k $(k=0,1,\dots,n)$ is a semisimple eigenvalue; and the associated eigenspace is spanned by functions $g_{\alpha,\beta}$'s with α and β satisfying $3|\alpha|+|\beta|=k$.

Proposition 3.6. *Let n be a nonnegative integer, and let r be an integer satisfying $r\geq n+3N+\frac{1}{2}$. Then for any fixed integers $\ell\geq 0$, $m\geq 0$, and $j=0,1$,*

$$\|\mathcal{Q}_nT(t)f_0\|_{X_r^{\ell,m+j}}\leq C(1+t^{-\frac{j}{2}})e^{\lambda_{n+1}t}\|f_0\|_{X_r^{\ell,m}}.$$

Proof. Let $r\geq n+3N+\frac{1}{2}$. By the Plancherel theorem it suffices to estimate $\|\partial_\xi^\alpha\partial_w^\beta(\xi^{\gamma_1}w^{\gamma_2}\widehat{\mathcal{Q}_nS(t)}f)\|_{L^2}$ for any α, β, γ_1, and γ_2 with $|\alpha|+|\beta|\leq r$.

We first make some preliminary observations. As in Proposition 3.5, we set $F(\tilde{\xi},\tilde{w}) = e^{\tilde{\mu}(\tilde{\xi},\tilde{w})}\hat{f}(\tilde{\xi},\tilde{w}-\tilde{\xi})$ and $\tilde{\mu}(\tilde{\xi},\tilde{w}) = |\tilde{w}-\frac{\tilde{\xi}}{2}|^2 + \frac{1}{12}|\tilde{\xi}|^2$. Then direct calculations give the bounds

$$\begin{aligned} &|\partial_{\tilde{\xi}}^{\alpha}\partial_{\tilde{w}}^{\beta}F(\tau c_1\tilde{\xi},\tau c_2\tilde{w})| \\ &\quad\le\ C(1+|\tilde{\xi}|+|\tilde{w}|)^{|\alpha|+|\beta|}e^{\tilde{\mu}(\tau c_1\tilde{\xi},\tau c_2\tilde{w})} \\ &\qquad\times\sum_{\substack{\sigma\le\alpha,\\ \eta'\le\beta}}\left|(\partial_\xi-\partial_w)^{\sigma}\partial_w^{\eta}\hat{f}(\tau c_1\tilde{\xi},\tau(c_2\tilde{w}-c_1\tilde{\xi}))\right| \end{aligned}\tag{3.1}$$

uniformly in τ, c_1, $c_2\in[0,1]$, and $\tilde{\xi}$, $\tilde{w}\in\mathbf{R}^N$, and

$$\hat{\mu}(\xi,w)-\tilde{\mu}(\tau e^{-\frac{3}{2}t}\xi,\tau e^{-\frac{1}{2}t}(w+\xi))\ge\frac{1}{2}\hat{\mu}(\xi,w)\tag{3.2}$$

for all $\tau\in[0,1]$, ξ, $w\in\mathbf{R}^N$, and sufficiently large t, e.g., $t\ge\log 36$.

We next set

$$F_n(\tilde{\xi},\tilde{w},t)=F(c_1\tilde{\xi},c_2\tilde{w})-\sum_{3|\alpha|+|\beta|\le n}\frac{1}{\alpha!\beta!}\partial_{\tilde{\xi}}^{\alpha}\partial_{\tilde{w}}^{\beta}F(0,0)c_1^{|\alpha|}\tilde{\xi}^{\alpha}c_2^{|\beta|}\tilde{w}^{\beta},$$

where $c_1=e^{-\frac{3}{2}t}$ and $c_2=e^{-\frac{1}{2}t}$.

Then if we set $\tilde{\xi}=\xi$ and $\tilde{w}=w+\xi$, Proposition 3.5 implies that

$$\begin{aligned} (\widehat{\mathcal{Q}_nS(t)}f)(\xi,w) &= (\widehat{S(t)}f)(\xi,w)-\sum_{k=0}^{n}(\widehat{P_kS(t)}f)(\xi,w) \\ &= e^{-\hat{\mu}(\tilde{\xi},\tilde{w}-\tilde{\xi})}F_n(\tilde{\xi},\tilde{w},t), \end{aligned}$$

and it suffices to estimate

$$\begin{aligned} &(\partial_\xi-\partial_w)^{\alpha'}\partial_w^{\beta'}\left[\xi^{\gamma_1}w^{\gamma_2}(\widehat{\mathcal{Q}_nS(t)}f)(\xi,w)\right] \\ &\quad=\partial_{\tilde{\xi}}^{\alpha'}\partial_{\tilde{w}}^{\beta'}\left[\tilde{\xi}^{\gamma_1}\tilde{w}^{\gamma_2}e^{-\hat{\mu}(\tilde{\xi},\tilde{w}-\tilde{\xi})}F_n(\tilde{\xi},\tilde{w},t)\right] \\ &\quad=\sum_{\tilde{\alpha}\le\alpha',\,\tilde{\beta}\le\beta'}\begin{pmatrix}\alpha'\\ \tilde{\alpha}\end{pmatrix}\begin{pmatrix}\beta'\\ \tilde{\beta}\end{pmatrix}\partial_{\tilde{\xi}}^{\alpha'-\tilde{\alpha}}\partial_{\tilde{w}}^{\beta'-\tilde{\beta}}\left[\tilde{\xi}^{\gamma_1}\tilde{w}^{\gamma_2}e^{-\hat{\mu}(\tilde{\xi},\tilde{w}-\tilde{\xi})}\right]\partial_{\tilde{\xi}}^{\tilde{\alpha}}\partial_{\tilde{w}}^{\tilde{\beta}}F_n(\tilde{\xi},\tilde{w},t) \\ &\quad\equiv\sum_{\tilde{\alpha}\le\alpha',\,\tilde{\beta}\le\beta'}I_{\alpha',\beta',\tilde{\alpha},\tilde{\beta},n} \end{aligned}$$

for all α' and β' with $|\alpha'|+|\beta'|\le r$.

In what follows we set $\tilde{\xi}=\xi$ and $\tilde{w}=w+\xi$.

Applying Taylor's theorem to $\partial_{\tilde{\xi}}^{\tilde{\alpha}}\partial_{\tilde{w}}^{\tilde{\beta}}F_n(\tilde{\xi},\tilde{w},t)$, we see that for multi-indices $\tilde{\alpha}$, $\tilde{\beta}$ with $|\tilde{\alpha}|+|\tilde{\beta}|=j\le n$,

$$\begin{aligned} &\partial_{\tilde{\xi}}^{\tilde{\alpha}}\partial_{\tilde{w}}^{\tilde{\beta}}F_n(\tilde{\xi},\tilde{w}) \\ =\ &\sum_{\substack{|\alpha|+|\beta|\\ =n-j+1}}c_1^{|\alpha|+|\tilde{\alpha}|}c_2^{|\beta|+|\tilde{\beta}|}\int_0^1\tfrac{(n-j+1)(1-\tau)^{n-j+1}}{\alpha!\beta!}\partial_{\tilde{\xi}}^{\alpha}\partial_{\tilde{w}}^{\beta}F(\tau c_1\tilde{\xi},\tau c_2\tilde{w})\,\tilde{\xi}^{\alpha}\tilde{w}^{\beta}\,d\tau \\ &+\sum_{(\alpha,\beta)\in\Lambda_{\tilde{\alpha},\tilde{\beta}}}\frac{1}{\tilde{\alpha}!\tilde{\beta}!}\partial_{\tilde{\xi}}^{\alpha}\partial_{\tilde{w}}^{\beta}F(0,0)c_1^{|\alpha|}\tilde{\xi}^{\alpha-\tilde{\alpha}}c_2^{|\beta|}\tilde{w}^{\beta-\tilde{\beta}} \\ \equiv\ &G^{(1)}_{\tilde{\alpha},\tilde{\beta},n}(\tilde{\xi},\tilde{w},t)+G^{(2)}_{\tilde{\alpha},\tilde{\beta},n}(\tilde{\xi},\tilde{w},t), \end{aligned}\tag{3.3}$$

where $\Lambda_{\tilde{\alpha},\tilde{\beta}} = \{(\alpha,\beta) : \alpha \geq \tilde{\alpha}, \beta \geq \tilde{\beta}, |\alpha| + |\beta| \leq n, 3|\alpha| + |\beta| \geq n+1\}$; $\tilde{\xi} = \xi$, $\tilde{w} = w + \xi$, $c_1 = e^{-\frac{3}{2}t}$, and $c_2 = e^{-\frac{1}{2}t}$. For $|\tilde{\alpha}| + |\tilde{\beta}| = j \geq n+1$, we have

$$\partial_{\tilde{\xi}}^{\tilde{\alpha}} \partial_{\tilde{w}}^{\tilde{\beta}} F_n(\tilde{\xi}, \tilde{w}) = c_1^{|\tilde{\alpha}|} c_2^{|\tilde{\beta}|} \partial_{\tilde{\xi}}^{\tilde{\alpha}} \partial_{\tilde{w}}^{\tilde{\beta}} F(c_1\tilde{\xi}, c_2\tilde{w}), \tag{3.4}$$

where $\tilde{\xi} = \xi$, $\tilde{w} = w + \xi$, $c_1 = e^{-\frac{3}{2}t}$, and $c_2 = e^{-\frac{1}{2}t}$.

Since $c_1^{|\alpha|} c_2^{|\beta|} = e^{-(\frac{3}{2}|\alpha| + \frac{1}{2}|\beta|)t} \leq e^{-\frac{n+1}{2}t}$ for $3|\alpha| + |\beta| \geq n+1$, we see from (3.1) and (3.3) that for $|\tilde{\alpha}| + |\tilde{\beta}| = j \leq n$,

$$\begin{aligned} & |G^{(2)}_{\tilde{\alpha},\tilde{\beta},n}(\tilde{\xi}, \tilde{w}, t)| \\ \leq\ & Ce^{-\frac{n+1}{2}t}(1 + |\xi| + |w|)^{n-j} \sup_{\sigma \leq \tilde{\alpha},\, \eta \leq \tilde{\beta}} \sup_{\xi, w} \left| (\partial_\xi - \partial_w)^\sigma \partial_w^\eta \hat{f}(\xi, w) \right| \\ \leq\ & Ce^{-\frac{n+1}{2}t}(1 + |\xi| + |w|)^{n-j} \int (1 + |x-u| + |u|)^n |f| \, dx du \\ \leq\ & Ce^{-\frac{n+1}{2}t}(1 + |\xi| + |w|)^{n-j} \|f\|_{X^{0,0}_{n+\frac{1}{2}(2N+1)}}. \end{aligned} \tag{3.5}$$

We also see from (3.1) and (3.3) that for $|\tilde{\alpha}| + |\tilde{\beta}| = j \leq n$,

$$\begin{aligned} & |G^{(1)}_{\tilde{\alpha},\tilde{\beta},n}(\tilde{\xi}, \tilde{w}, t)| \\ \leq\ & C(1 + |\xi| + |w|)^{2(n+1)} \left(\int_0^1 e^{\tilde{\mu}(\tau c_1 \tilde{\xi}, \tau c_2 \tilde{w})} \, d\tau \right) \\ & \times \sum_{\substack{|\alpha|+|\beta| \\ = n-j+1}} e^{-\{\frac{3}{2}(|\alpha| + |\tilde{\alpha}|) + \frac{1}{2}(|\beta| + |\tilde{\beta}|)\}t} \sup_{\substack{\sigma \leq \alpha,\, \eta \leq \beta \\ \xi, w}} \left| (\partial_\xi - \partial_w)^\sigma \partial_w^\eta \hat{f}(\xi, w) \right| \\ \leq\ & Ce^{-\frac{n+1}{2}t} \|f\|_{X^{0,0}_{n+1+\frac{1}{2}(2N+1)}} (1 + |\xi| + |w|)^{2(n+1)} \int_0^1 e^{\tilde{\mu}(\tau c_1 \tilde{\xi}, \tau c_2 \tilde{w})} \, d\tau. \end{aligned} \tag{3.6}$$

It then follows from (3.2), (3.5), and (3.6) that for $|\tilde{\alpha}| + |\tilde{\beta}| = j \leq n$

$$|I_{\alpha',\beta',\tilde{\alpha},\tilde{\beta},n}| \leq Ce^{-\frac{n+1}{2}t} \|f\|_{X^{0,0}_r} e^{-\frac{1}{4}\hat{\mu}(\xi,w)},$$

provided that $t \geq \log 36$ since $r \geq n + 3N + \frac{1}{2} > n + 1 + \frac{1}{2}(2N+1)$. This implies that for $|\tilde{\alpha}| + |\tilde{\beta}| = j \leq n$

$$\|I_{\alpha',\beta',\tilde{\alpha},\tilde{\beta},n}\|_{L^2} \leq Ce^{-\frac{n+1}{2}t} \|f\|_{X^{0,0}_r},$$

provided that $t \geq \log 36$.

For $j = |\tilde{\alpha}| + |\tilde{\beta}| \geq n+1$, we apply (3.1), (3.2), and (3.4) to obtain

$$\begin{aligned} |I_{\alpha',\beta',\tilde{\alpha},\tilde{\beta},n}| \leq\ & Ce^{-(\frac{3}{2}|\tilde{\alpha}| + \frac{1}{2}|\tilde{\beta}|)t} e^{-\frac{1}{4}\hat{\mu}(\xi,w)} \\ & \times \sum_{|\sigma| + |\eta| \leq j} \left| (\partial_\xi - \partial_w)^\sigma \partial_w^\eta \hat{f}(e^{-\frac{3}{2}t}\xi, e^{-\frac{1}{2}t}(w + \xi) - e^{-\frac{3}{2}t}\xi) \right| \end{aligned}$$

for $t \geq \log 36$. Thus we have, for $t \geq \log 36$ and $j = |\tilde{\alpha}| + |\tilde{\beta}| \geq n+1$,

$$\|I_{\alpha',\beta',\tilde{\alpha},\tilde{\beta},n}\|_{L^2} \leq C \left\{ \begin{array}{l} e^{-(\frac{3}{2}|\tilde{\alpha}| + \frac{1}{2}|\tilde{\beta}|)t} \|f\|_{X^{0,0}_{j+\frac{1}{2}(2N+1)}}, \\ e^{Nt - (\frac{3}{2}|\tilde{\alpha}| + \frac{1}{2}|\tilde{\beta}|)t} \|f\|_{X^{0,0}_j}. \end{array} \right. \tag{3.7}$$

For $n+2N+1 \le j = |\tilde{\alpha}|+|\tilde{\beta}| \le r$ we apply the second inequality of (3.7) to obtain

$$\|I_{\alpha',\beta',\tilde{\alpha},\tilde{\beta},n}\|_{L^2} \le Ce^{-\frac{n+1}{2}t}\|f\|_{X_r^{0,0}},$$

since $N - (\frac{3}{2}|\tilde{\alpha}| + \frac{1}{2}|\tilde{\beta}|) \le -\frac{n+1}{2}$; while for $n+1 \le j = |\tilde{\alpha}|+|\tilde{\beta}| \le n+2N$ we apply the first inequality of (3.7) to obtain

$$\|I_{\alpha',\beta',\tilde{\alpha},\tilde{\beta},n}\|_{L^2} \le Ce^{-\frac{n+1}{2}t}\|f\|_{X_r^{0,0}},$$

since $j + \frac{1}{2}(2N+1) \le n+3N+\frac{1}{2} \le r$.

Now recall that $T(t) = e^{-(2N-\gamma)t}S(t)$. Then, for any nonnegative integers ℓ and m, we find from the above estimates that

$$\|\mathcal{Q}_n T(t)f\|_{X_r^{\ell,m}} \le Ce^{\lambda_{n+1}t}\|f\|_{X_r^{0,0}}$$

holds provided that $t \ge \log 36$.

For $t \le \log 36$, it is easy to see that

$$\|T(t)f\|_{X_r^{\ell,m+j}} \le C(1+t^{-\frac{j}{2}})\|f\|_{X_r^{\ell,m}}$$

for any nonnegative r, l, and m and $j=0,1$. Thus we have

$$\|\mathcal{Q}_n T(t)f\|_{X_r^{\ell,m+j}} \le C(1+t^{-\frac{j}{2}})\|f\|_{X_r^{\ell,m}}$$

for $t \le \log 36$. This, together with the estimate for $t \ge \log 36$, yields the desired result, and the proof is complete. □

Remark 3.7. One can also obtain

$$\|T(t)f\|_{X_r^{\ell+1,m}} \le C(1+t^{-3/2})\|f\|_{X_r^{\ell,m}}$$

for small t.

4. Estimates for the nonlinearity. Theorem 2.1 follows from Propositions 3.6 and 4.1 below as in [10, 21] by applying standard arguments of invariant manifold theory in [6, 16]. So our remaining task is to prove the following.

PROPOSITION 4.1. *Let $N \ge 2$. If $r > \frac{N}{2}$, then the following estimate holds for any fixed nonnegative m and k:*

$$\|E(f)\cdot\nabla_u g\|_{X_r^{m+\ell,k}} \le C\|f\|_{X_r^{m+\ell,k}}\|g\|_{X_r^{m+\ell,k+1}},$$

where $\ell = [\frac{N}{2}-1]+1$.

Proof. Here we prove the case $m=k=0$ only. The extension to general m and k is an easy task.

First, we observe that $\frac{x}{|x|^N}$ is homogeneous of degree $1-N$. Therefore,

$$\|\partial_x^\alpha E(f)\|_{L^q(dx)} \le C\|\int \partial_x^\alpha f(x,u)\,du\|_{L^p(dx)}$$

with $1<p<q<\infty$, $\frac{1}{q} = \frac{1}{p} - \frac{1}{N}$ [13, Cor. 5.15, pp. 137]. We thus obtain

$$\begin{aligned} \|\partial_x^\alpha E(f)\|_{L^q(dx)} &\le C\int \|\partial_x^\alpha f(x,u)\|_{L^p(dx)}\,du \\ &\le C_\varepsilon\left(\int (1+|u|^2)^{\frac{N+\varepsilon}{2}}\|\partial_x^\alpha f(x,u)\|_{L^p(dx)}^2\,du\right)^{1/2} \end{aligned} \tag{4.1}$$

for any $\varepsilon > 0$, where p and q are the same as above.

Second, we note that $\|f\|_{X_r^{\ell,0}}$ is equivalent with $\|\sum_{|\alpha|\leq\ell}\partial_x^\alpha(\rho_r f)\|_{L^2(dxdu)}$, where $\rho_r(x,u)=(1+|x|^2+|u|^2)^{r/2}$. Therefore, by the interpolation inequality,

$$\|\nabla_x^j f\|_{L^2(dx)}\leq C\|\nabla_x^\ell f\|_{L^2(dx)}^{\frac{j}{\ell}}\|f\|_{L^2(dx)}^{1-\frac{j}{\ell}},\quad 1\leq j\leq\ell,$$

it suffices to show that

$$\|\partial_x^\alpha(\rho_r E(f)\cdot\nabla_u g)\|_{L^2(dxdu)}\leq C\|f\|_{X_r^{\ell,0}}\|g\|_{X_r^{\ell,1}}$$

for $|\alpha|=\ell$ and 0.

For a multi-index β and $r\geq 0$ we set $\tilde g_{\beta,r}(x,u)=|\partial_x^\beta(\rho_r(x,u)\nabla_u g(x,u))|$. Then

$$\|\partial_x^\alpha(\rho_r E(f)\cdot\nabla_u g)\|_{L^2(dxdu)}\leq C\sum_{\beta\leq\alpha}\|\partial_x^{\alpha-\beta}E(f)\tilde g_{\beta,r}\|_{L^2(dxdu)}\equiv C\sum_{\beta\leq\alpha}J_{\alpha,\beta,r}$$

and

$$J_{\alpha,\beta,r}\leq\|\partial_x^{\alpha-\beta}E(f)\|_{L^q(dx)}\big\|\|\tilde g_{\beta,r}\|_{L^s(dx)}\big\|_{L^2(du)}\tag{4.2}$$

with $2\leq q,\ s\leq\infty$ satisfying $\frac{1}{q}+\frac{1}{s}=\frac{1}{2}$.

We consider the case $N=3$. Then $\ell=[\frac{N}{2}-1]+1=1$. We estimate each $J_{\alpha,\beta,r}$. For $|\alpha|=\ell(=1)$ and $|\beta|=0$ we take $\frac{1}{q}=\frac{1}{2}-\frac{1}{N}$ and $\frac{1}{s}=\frac{1}{N}$ in (4.2). Then by (4.1) with $p=2$ we have

$$\|\partial_x^\alpha E(f)\|_{L^q(dx)}\leq C_\varepsilon\|\partial_x^\alpha f\|_{X^{0,0}_{\frac{1}{2}(N+\varepsilon)}},$$

and by the Gagliardo–Nirenberg–Sobolev inequality (see, e.g., [14]), we see that

$$\|\tilde g_{\beta,r}\|_{L^s(dx)}\leq C\|\nabla_x^\ell\tilde g_{0,r}\|_{L^2(dx)}^{\delta}\|\tilde g_{0,r}\|_{L^2(dx)}^{1-\delta},$$

where $\frac{1}{s}=\delta(\frac{1}{2}-\frac{\ell}{N})+(1-\delta)\frac{1}{2}$ and $\delta\in[0,1]$. It then follows that for $|\alpha|=\ell(=1)$ and $|\beta|=0$

$$J_{\alpha,\beta,r}\leq C_\varepsilon\|f\|_{X^{\ell,0}_{\frac{1}{2}(N+\varepsilon)}}\|g\|_{X_r^{\ell,1}}.$$

We next estimate $J_{\alpha,\beta,r}$ with $|\beta|=\ell(=1)$. Note that in this case $\alpha=\beta$. Taking $q=\infty$ and $s=2$ in (4.2) and using (4.1) and the Gagliardo–Nirenberg–Sobolev inequality, we obtain

$$\|E(f)\|_{L^\infty(dx)}\leq C\|\nabla_x^\ell E(f)\|_{L^{q_1}(dx)}^{\delta}\|E(f)\|_{L^{q_1}(dx)}^{1-\delta}\leq C_\varepsilon\|f\|_{X^{\ell,0}_{\frac{1}{2}(N+\varepsilon)}},$$

where $\frac{1}{q_1}=\frac{1}{2}-\frac{1}{N}$, $0=\delta(\frac{1}{q_1}-\frac{\ell}{N})+(1-\delta)\frac{1}{q_1}$, and $\delta\in[0,1]$. Whence, for $|\alpha|=|\beta|=\ell(=1)$,

$$J_{\alpha,\beta,r}\leq C_\varepsilon\|f\|_{X^{\ell,0}_{\frac{1}{2}(N+\varepsilon)}}\|g\|_{X_r^{\ell,1}}.$$

In the case when $\alpha=0$, we take $q=\infty$ and $s=2$ in (4.2). Then, similarly as above, we see that

$$J_{\alpha,\beta,r}\leq C_\varepsilon\|f\|_{X^{\ell,0}_{\frac{1}{2}(N+\varepsilon)}}\|g\|_{X_r^{0,1}}$$

for $\alpha = \beta = 0$. Combining the above estimates, we obtain the desired estimate for $N = 3$.

We next consider the case when $N = 2$. In this case we also have $\ell = [\frac{N}{2} - 1] + 1 = 1$. We choose $1 < \tilde{p} < 2$ and set $\frac{1}{\tilde{q}} = \frac{1}{\tilde{p}} - \frac{1}{2}$. When $|\beta| = \ell (= 1)$, we take $q = \infty$ and $s = 2$ in (4.2). Then by (4.1) and the Gagliardo–Nirenberg–Sobolev inequality, we see that

$$\begin{aligned} \|E(f)\|_{L^\infty(dx)} &\leq C\|\nabla_x^\ell E(f)\|_{L^{\tilde{q}}(dx)}^\delta \|E(f)\|_{L^{\tilde{q}}(dx)}^{1-\delta} \\ &\leq \left(\sum_{|\alpha| \leq \ell} \int (1 + |u|^2)^{\frac{1}{2}(N+\varepsilon)} \|\partial_x^\alpha f\|_{L^{\tilde{p}}(dx)}^2 \, du \right)^{1/2} \\ &\leq C_\varepsilon \|f\|_{X^{\ell,0}_{\frac{1}{\tilde{p}}(N+\varepsilon)}}, \end{aligned}$$

where $0 = \delta(\frac{1}{\tilde{q}} - \frac{\ell}{N}) + (1 - \delta)\frac{1}{\tilde{q}}$ and $\delta \in [0, 1]$. Therefore, for $|\alpha| = |\beta| = \ell (= 1)$,

$$J_{\alpha,\beta,r} \leq C_\varepsilon \|f\|_{X^{\ell,0}_{\frac{1}{\tilde{p}}(N+\varepsilon)}} \|g\|_{X_r^{\ell,1}}.$$

For $\alpha = \beta = 0$ we obtain, in a similar manner,

$$J_{\alpha,\beta,r} \leq C_\varepsilon \|f\|_{X^{\ell,0}_{\frac{1}{\tilde{p}}(N+\varepsilon)}} \|g\|_{X_r^{0,1}}.$$

For $|\alpha| = \ell (= 1)$ and $|\beta| = 0$ we take $q = \tilde{q}$ and $\frac{1}{s} = \frac{1}{2} - \frac{1}{\tilde{q}}$ in (4.2). By (4.1) we have

$$\|\partial_x^\alpha E(f)\|_{L^{\tilde{q}}(dx)} \leq C_\varepsilon \|\partial_x^\alpha f\|_{X^{0,0}_{\frac{1}{\tilde{p}}(N+\varepsilon)}}.$$

By the Gagliardo–Nirenberg–Sobolev inequality, we have

$$\|\tilde{g}_{\beta,r}\|_{L^s(dx)} \leq C\|\nabla_x^\ell \tilde{g}_{0,r}\|_{L^2(dx)}^\delta \|\tilde{g}_{0,r}\|_{L^2(dx)}^{1-\delta},$$

where $\frac{1}{s} = \frac{1}{2} - \frac{1}{\tilde{q}}$, $\frac{1}{s} = \delta(\frac{1}{2} - \frac{\ell}{N}) + (1 - \delta)\frac{1}{2}$, and $\delta \in [0, 1]$. It then follows that for $|\alpha| = \ell (= 1)$ and $|\beta| = 0$

$$J_{\alpha,\beta,r} \leq C_\varepsilon \|f\|_{X^{\ell,0}_{\frac{1}{\tilde{p}}(N+\varepsilon)}} \|g\|_{X_r^{\ell,1}}.$$

Combining the above estimates, we obtain

$$\|E(f) \cdot \nabla_u g\|_{X_r^{\ell,0}} \leq C_{\varepsilon,\tilde{p}} \|f\|_{X^{\ell,0}_{\frac{1}{\tilde{p}}(N+\varepsilon)}} \|g\|_{X_r^{\ell,1}}.$$

Now, for a given $r > \frac{N}{2}$, choose $\tilde{p}$ and $\varepsilon > 0$ so that $r > \frac{1}{\tilde{p}}(N + \varepsilon)$. Then we obtain the desired estimate for the case when $N = 2$.

When N is odd and $N \geq 5$, $J_{\alpha,\beta,r}$ is estimated as follows. If $|\alpha| = 0$ or if $|\alpha| = \ell$ and $|\beta| = 0$, ℓ, then one can estimate $J_{\alpha,\beta,r}$ in the same way as in the case when $N = 3$ since $\ell = [\frac{N}{2} - 1] + 1$. For $|\alpha| = \ell$ and $1 \leq |\beta| \leq \ell - 1$, we can apply the Gagliardo–Nirenberg–Sobolev inequality to obtain

$$\|\partial_x^{\alpha-\beta} E(f)\|_{L^q(dx)} \leq C\|\nabla_x^\ell E(f)\|_{L^{q_1}(dx)}^{\delta_1} \|E(f)\|_{L^{q_1}(dx)}^{1-\delta_1}$$

and

$$\|\tilde{g}_{\beta,r}\|_{L^s(dx)} \leq C\|\nabla_x^\ell \tilde{g}_{0,r}\|_{L^2(dx)}^{\delta_2} \|\tilde{g}_{0,r}\|_{L^2(dx)}^{1-\delta_2}$$

for some $2 < q,\, q_1,\, s < \infty$, $\frac{\ell-|\beta|}{\ell} \le \delta_1 < 1$, and $\frac{|\beta|}{\ell} \le \delta_2 < 1$ with

$$\text{(4.3)} \qquad \begin{cases} \dfrac{1}{q} = \dfrac{\ell-|\beta|}{N} + \delta_1\left(\dfrac{1}{q_1} - \dfrac{\ell}{N}\right) + (1-\delta_1)\dfrac{1}{q_1}, \quad \dfrac{1}{q_1} = \dfrac{1}{2} - \dfrac{1}{N}, \\ \dfrac{1}{s} = \dfrac{|\beta|}{N} + \delta_2\left(\dfrac{1}{2} - \dfrac{\ell}{N}\right) + (1-\delta_2)\dfrac{1}{2}, \quad \dfrac{1}{s} = \dfrac{1}{2} - \dfrac{1}{q}. \end{cases}$$

In fact, this is possible since $\ell = [\frac{N}{2} - 1] + 1$. For example, take $\delta_1 = \frac{N-\frac{3}{2}}{N-1}$ in (4.3). We then verify that $2 < q < \infty$ and $\delta_2 = \delta_1$. Using (4.1), we can now obtain the desired estimate.

When N is even and $N \ge 4$, one can prove similarly to the case of odd N with a slight modification as we did for $N = 2$. Set q_1 and δ_1 in (4.3) such as $\frac{1}{q_1} = \frac{1}{\tilde{p}} - \frac{1}{N}$ and $\delta_1 = \frac{1}{\tilde{p}} - \frac{1}{2} + \frac{N-1}{N}$ for $1 < \tilde{p} < 2$. If $\tilde{p}$ is sufficiently close to 2, then we find that $2 < q < \infty$ and $\frac{\ell-|\beta|}{\ell} \le \delta_1 < 1$. The desired estimate then follows in a similar manner to the case when $N = 2$. This completes the proof. □

Appendix. In the appendix we outline a derivation of an integral formula of $S(t)f_0$.

Consider the linear problem

$$\text{(A.1)} \qquad \begin{aligned} &\partial_t f - \left(\frac{3}{2}x - u\right)\cdot\nabla_x f - \frac{1}{2}u\cdot\nabla_u f - 2Nf - \Delta_u f = 0, \\ &f|_{t=0} = f_0. \end{aligned}$$

Taking the Fourier transform of (A.1), we have

$$\text{(A.2)} \qquad \begin{aligned} &\partial_t \hat{f} + \frac{3}{2}\xi\cdot\nabla_\xi \hat{f} + \left(\frac{1}{2}w - \xi\right)\cdot\nabla_w \hat{f} + |w|^2\hat{f} = 0, \\ &\hat{f}|_{t=0} = \hat{f}_0. \end{aligned}$$

The problem (A.2) is reduced to the following two problems:

$$\text{(A.3)} \qquad \partial_t \hat{f} + \frac{3}{2}\xi\cdot\nabla_\xi \hat{f} + \left(\frac{1}{2}w - \xi\right)\cdot\nabla_w \hat{f} = 0, \quad \hat{f}|_{t=0} = \hat{f}_0,$$

and

$$\text{(A.4)} \qquad \partial_t \hat{f} + \frac{3}{2}\xi\cdot\nabla_\xi \hat{f} + \left(\frac{1}{2}w - \xi\right)\cdot\nabla_w \hat{f} = -|w|^2, \quad \hat{f}|_{t=0} = 0.$$

Let $\hat{f}^{(1)}(\xi, w, t)$ and $\hat{f}^{(2)}(\xi, w, t)$ be solutions of (A.3) and (A.4), respectively. Then the solution of (A.2) is given by

$$\hat{f}(\xi, w, t) = \hat{f}^{(1)}(\xi, w, t)\, e^{\hat{f}^{(2)}(\xi, w, t)}.$$

Solutions of (A.3) and (A.4) can be obtained easily by the characteristics. Let $(\Xi(t), W(t))$ be the solution of

$$\frac{d}{dt}\begin{pmatrix} \Xi \\ W \end{pmatrix} = \begin{pmatrix} -\frac{3}{2} & 0 \\ 1 & -\frac{1}{2} \end{pmatrix}\begin{pmatrix} \Xi \\ W \end{pmatrix}, \quad \begin{pmatrix} \Xi(0) \\ W(0) \end{pmatrix} = \begin{pmatrix} \xi \\ w \end{pmatrix},$$

which are explicitly given as

$$\Xi(t) = e^{-\frac{3}{2}t}\xi, \quad W(t) = e^{-\frac{1}{2}t}w + (e^{-\frac{1}{2}t} - e^{-\frac{3}{2}t})\xi.$$

Then solutions $\hat{f}^{(1)}(\xi, w, t)$ and $\hat{f}^{(2)}(\xi, w, t)$ take the forms

$$\hat{f}^{(1)}(\xi, w, t) = \hat{f}_0(\Xi(t), W(t)) = \hat{f}_0(e^{-\frac{3}{2}t}\xi, e^{-\frac{1}{2}t}w + e^{-\frac{1}{2}t}a(t)\xi)$$

and

$$\hat{f}^{(2)}(\xi, w, t) = -\int_0^t |W(t-s)|^2\, ds = -a(t)\left|w + \frac{a(t)}{2}\xi\right|^2 - \frac{a(t)^3}{12}|\xi|^2,$$

where $a(t) = 1 - e^{-t}$. Hence the solution of (A.2) is

$$\widehat{S(t)f_0}(\xi, w) = \hat{f}(\xi, w, t) = \hat{f}_0(e^{-\frac{3}{2}t}\xi, e^{-\frac{1}{2}t}w + e^{-\frac{1}{2}t}a(t)\xi)e^{-a(t)|w+\frac{a(t)}{2}\xi|^2 - \frac{a(t)^3}{12}|\xi|^2}.$$

The inverse Fourier transform now gives an integral formula of the semigroup $S(t)f_0$:

$$S(t)f_0 = \left(\frac{\sqrt{3}}{2\pi}\right)^N a(t)^{-2N} \int e^{-\frac{3|x-e^{-\frac{3}{2}t}y - \frac{a(t)}{2}(u+e^{-\frac{t}{2}}v)|^2}{a(t)^3} - \frac{|u-e^{-\frac{t}{2}}v|^2}{4a(t)}} f_0(y, v)dydv,$$

where $a(t) = 1 - e^{-t}$.

Remarks. (i) One can see that the semigroup $S(t)$ is compact for $t > 0$ in the space $X = L^2(e^{\mu(x,u)}dxdu)$ with $\mu(x, u) = 3|x - \frac{u}{2}|^2 + \frac{|u|^2}{4}$ (which implies that the spectrum of $S(t)$ consists only of discrete eigenvalues). This can be shown as follows. A straightforward calculation yields the identity

$$\begin{aligned} &\frac{3\left|x - e^{-\frac{3}{2}t}y - \dfrac{a(t)}{2}(u + e^{-\frac{t}{2}}v)\right|^2}{a(t)^3} + \frac{|u - e^{-\frac{t}{2}}v|^2}{4a(t)} + \frac{1}{2}\mu(y, v) - \frac{1}{2}\mu(x, u) \\ &= \frac{\nu_1(a(t))}{a(t)^3}\left|x - \frac{F_1(t, a(t), y, u, v)}{G_1(a(t))}\right|^2 + \frac{\nu_2(a(t))}{a(t)}\left|u - \frac{F_2(t, a(t), y, v)}{G_2(a(t))}\right|^2 \\ &\quad + a(t)\nu_3(a(t))|y - F_3(t, a(t), v)|^2 + a(t)^3\nu_4(a(t))|v|^2, \end{aligned}$$

where $\nu_j(a)$ $(j = 1, \dots, 4)$ are smooth in a with $\delta \le \nu_j(a) \le M$ for some $\delta > 0$ and $M > 0$ uniformly in $a \in [0, 1]$, F_j are polynomials in a whose coefficients are smooth in t and linear in y, u, v, and G_j are polynomials in a with $\inf_{0\le a\le 1} G_j(a) \ge \delta$ for some $\delta > 0$. Noting this identity and using the integral formula of $S(t)f_0$ above, one can see that

$$\|(1 + \nabla_x + \nabla_u)S(t)f_0\|_X \le Ce^{\eta t}(1 + a(t)^{-3/2} + a(t)^{-1/2})\|f_0\|_X$$

for $t > 0$ with some constants $\eta > 0$ and $C > 0$. This inequality shows that $S(t)$ is compact for $t > 0$ in X, since the embedding

$$\{f \in X : \nabla_x f, \nabla_u f \in X\} \hookrightarrow X$$

is compact [12, Proposition 1.1].

(ii) Noting remark (i) and the fact that $\hat{f}(\xi, w)$ is analytic in ξ and w for $f \in X$, one can deduce that the spectrum $\sigma(A)$ of A in X consists only of discrete eigenvalues $\sigma_k = -\frac{k}{2}$, $k = 0, 1, 2, \dots$, and the eigenspace associated with σ_k is spanned by functions $g_{\alpha,\beta}$ with $3|\alpha| + |\beta| = k$. The corresponding assertion holds for the adjoint A^*.

(iii) As for $\frac{d}{dt}S(t)$, the inequality

$$\left\| \frac{d}{dt} S(t) f_0 \right\|_X \leq C t^{-2} \|f_0\|_X$$

holds for small $t > 0$, and the behavior $\|\frac{d}{dt}S(t)f_0\|_X = O(t^{-2})$ as $t \to 0$ seems to be optimal. Thus $S(t)$ does not seem to be analytic in X.

Acknowledgments. The author is very grateful to the referees for their valuable comments. The author expresses his hearty thanks to Professor Takayuki Kobayashi for valuable discussions. The author also thanks Professor Takayoshi Ogawa for his interest in this work.

REFERENCES

[1] F. Bouchut, *Existence and uniqueness of a global smooth solution for the Vlasov-Poisson-Fokker-Planck system in three dimensions*, J. Funct. Anal., 111 (1993), pp. 239–258.

[2] F. Bouchut, *Smoothing effect for the non-linear Vlasov-Poisson-Fokker-Planck system*, J. Differential Equations, 122 (1995), pp. 225–238.

[3] F. Bouchut and J. Dolbeault, *On long time asymptotics of the Vlasov-Fokker-Planck equation and of the Vlasov-Poisson-Fokker-Planck system with Coulombic and Newtonian potentials*, Differential Integral Equations, 8 (1995), pp. 487–514.

[4] J. Bricmont, A. Kupiainen, and G. Lin, *Renormalization group and asymptotics of solutions of nonlinear parabolic equations*, Comm. Pure Appl. Math., 47 (1994), pp. 893–922.

[5] A. Carpio, *Long-time behaviour for solutions of the Vlasov-Poisson-Fokker-Planck equation*, Math. Methods Appl. Sci., 21 (1998), pp. 985–1014.

[6] J. Carr, *Applications of Center Manifold Theory*, Springer-Verlag, New York, Heidelberg, Berlin, 1981.

[7] J. A. Carrillo and J. Soler, *On the initial value problem for the Vlassov-Poisson-Fokker-Planck system with initial data in L^p spaces*, Math. Methods Appl. Sci., 18 (1995), pp. 825–839.

[8] J. A. Carrillo and J. Soler, *On the Vlasov-Poisson-Fokker-Planck equations with measures in Morrey spaces as initial data*, J. Math. Anal. Appl., 207 (1997), pp. 475–495.

[9] J. A. Carrillo, J. Soler, and J. L. Vázquez, *Asymptotic behaviour and self-similarity for the three dimensional Vlasov-Poisson-Fokker-Planck system*, J. Funct. Anal., 141 (1996), pp. 99–132.

[10] J.-P. Eckmann and C. E. Wayne, *Non-linear stability analysis of higher order dissipative partial differential equations*, Math. Phys. Electron. J., 4 (1998), Paper 3, 20 pp.

[11] J.-P. Eckmann, C. E. Wayne, and P. Wittwer, *Geometric stability analysis for periodic solutions of the Swift-Hohenberg equation*, Comm. Math. Phys., 190 (1997), pp. 173–211.

[12] M. Escobedo and O. Kavian, *Variational problems related to self-similar solutions of the heat equation*, Nonlinear Anal., 11 (1987), pp. 1103–1133.

[13] G. B. Folland, *Lectures on Partial Differential Equations*, Tata Institute of Fundamental Research Lectures on Mathematics and Physics, Bombay, 1983.

[14] A. Friedman, *Partial Differential Equations*, Holt, Rinehart, and Winston, New York, 1969.

[15] K. Kobayashi, *An L^p Theory of Invariant Manifolds for Parabolic Partial Differential Equations on R^d*, preprint, Waseda University, Tokyo, 2000.

[16] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, Heidelberg, 1981.

[17] K. Ono and W. Strauss, *Regular solutions of the Vlasov-Poisson-Fokker-Planck system*, Discrete Contin. Dynam. Systems, 6 (2000), pp. 751–772.

[18] G. Rein and J. Weckler, *Generic global classical solutions of the Vlasov-Fokker-Planck-Poisson system in three dimensions*, J. Differential Equations, 99 (1992), pp. 59–77.

[19] H. D. Victory, *On the existence of global weak solutions for Vlasov-Poisson-Fokker-Planck systems*, J. Math. Anal. Appl., 160 (1991), pp. 525–555.

[20] H. D. Victory and B. P. O'Dwyer, *On classical solutions of Vlasov-Poisson Fokker-Planck systems*, Indiana Univ. Math. J., 39 (1990), pp. 105–156.

[21] C. E. Wayne, *Invariant manifolds for parabolic partial differential equations on unbounded domains*, Arch. Rational Mech. Anal., 138 (1997), pp. 279–306.

[22] C. E. Wayne, *Invariant manifolds and the asymptotics of parabolic equations in cylindrical domains*, in Differential Equations and Applications (Hagzhou, 1996), P. W. Bates, S.-N. Chow, K. Lu, and X. Pan, eds., International Press, Cambridge, MA, 1997, pp. 314–325.

SIAM J. MATH. ANAL.
Vol. 33, No. 3, pp. 509–522

ON THE UNIQUENESS OF THE CONTINUATION FOR A THERMOELASTICITY SYSTEM[*]

VICTOR ISAKOV[†]

Abstract. We obtain new uniqueness of the continuation results for the thermoelasticity system on the plane. The crucial ingredient of the proofs is the use of Carleman-type estimates with two large parameters for basic second order partial differential operators with constant coefficients. We derive these estimates by applying differential quadratic forms. The proposed technique can be of value when studying similar questions for systems of partial differential equations of upper triangular principal structure. The results can be applied to control theory and inverse problems.

Key words. uniqueness of the continuation, Carleman estimates, the equations of thermoelasticity

AMS subject classifications. 35Bxx, 35Lxx

PII. S0036141000366509

We will consider the lateral Cauchy problem for a thermoelasticity system (1.1), (1.2). This system describes combined elastic and thermal effects in plates and is of current theoretical and applied interest. We observe that this system is strongly coupled and has an "upper triangular" principal structure. Many important questions including the fundamental one about uniqueness of the continuation are open. The system (1.1), (1.2) cannot be principally diagonalized like Maxwell's or elasticity systems in the recent work of Eller, Isakov, Nakamura, and Tataru [4] and Isakov [9]. In this paper we give sufficient conditions for uniqueness by combining new Carleman estimates with additional large parameter λ for the Laplace, wave, and heat equations. This additional parameter enables us to principally decouple the system. For wave operators such estimates were first introduced and used for the elasticity system in the paper of Isakov [7]. In section 3 of this paper we derive them with the same weight function for the wave and Laplace equations, and in section 4 we derive them for the heat operator. Finally, in section 5 we combine these estimates to obtain the main uniqueness result. In case of constant (or more generally, analytic) coefficients one can derive sharp uniqueness results from the Holmgren–John theorem, and we give this short derivation in section 2. But when coefficients are of finite smoothness, this argument cannot be applied as well as its recent generalization by Tataru [16] and the subsequent work of Robbiano and Zuily [15]. The paper of Lebeau and Zuazua [13] handles a thermoelasticity system in self-adjoint form (with variable principal part) because it requires use of eigenfunctions expansions and of the geometrical optics technique introduced by Bardos, Lebeau, and Rauch [3] with its restrictions of extra regularity of coefficients and absence of diffractive points of the lateral boundary.

We expect that the idea of the additional parameter λ can be used for other systems with "triangular" principal structure. Observe that the concept of the principal part depends on a particular equation or system. Some interesting cases are discussed in the book of Hörmander [5] as well as in the papers of Isakov [8] and Nirenberg [14].

[*] Received by the editors January 14, 2000; accepted for publication (in revised form) September 20, 2000; published electronically September 5, 2001. This research was supported in part by NSF grants DMS 95-01510 and DMS-9803397.

http://www.siam.org/journals/sima/33-3/36650.html

[†] Department of Mathematics and Statistics, Wichita State University, Wichita, KS 67260-0033 (Victor.Isakov@wichita.edu).

Uniqueness of the continuation has immediate applications to inverse problems (see Isakov [10]) and to optimal control theory. In particular, by using uniqueness of the continuation, Avalos and Lasiecka [2] recently proved boundary stabilizability of the system (1.1), (1.2). The results of Lagnese [12] about exact-approximate controllability are applicable only to the system (1.1), (1.2) with constant coefficients when control is applied to the whole lateral boundary, and there is a good possibility that the results of this paper and their extensions to variable coefficients will be missing ingredients for proofs of controllability in a general situation, in particular from a part of the lateral boundary.

We use the standard notation $\partial_k = \partial/\partial x_k$, $\partial_t = \partial/\partial_t$, $D = -i\partial$. We let $\nabla = (\partial_1, \ldots, \partial_n)$, $\nabla_{x,t} = (\nabla, \partial_t)$. α and β are multi-indices of partial differentiations ∂^α, ∂^β. $B(a;r)$ is the ball of radius r centered at a. ν denotes the exterior unit normal to the boundary of a domain, C and δ denote constants depending only on operators P, the domain Q, the weight function ϕ, and the parameter $\epsilon > 0$. We will specify dependence on all other parameters. $\| \|_{(k)}(Q)$ is the norm in the Sobolev space $H_k(Q)$.

1. Main results. We are interested in the system of two strongly coupled partial differential equations

$$\partial_t^2 w - \gamma\Delta\partial_t^2 w + \Delta^2 w + b_1\Delta v + B_1 w + B_2 v = 0 \quad \text{in } Q = \Omega \times (0,T), \tag{1.1}$$

$$\partial_t v - b\Delta v + B_3 w + B_4 v = 0 \qquad \text{in } Q = \Omega \times (0,T), \tag{1.2}$$

where γ, b are positive constants, Ω is a domain in $\mathbb{R}^2$,

$$B_k w = b_k \partial_t \Delta w + \sum b_{j,k}\partial_j \Delta w + \sum b_{2+j,k}\partial_t^2 \partial_j w + \sum b_{k,\alpha}\partial^\alpha w, k = 1, 3,$$

and the sums in B_k are over $j = 1, 2, |\alpha| \leq 2$, with $\alpha_0 = 0$ for B_4, $B_k v = \sum b_{4+j,k}\partial_j v + b_{7,k} v$, $k = 2, 4$. Here b_k, $b_{j,k}$ are in $L^\infty(Q)$. In particular, when $B_1 = B_2 = 0$, $B_3 = b_3\partial_t\Delta$, $B_4 v = b_{7,4} v$, we obtain the simplest thermoelasticity system for (scalar) displacement w and temperature v. We are interested in the (lateral) Cauchy problem for this system when in addition to the equations we prescribe the Cauchy data

$$\partial_\nu^j w = w_j, \quad \partial_\nu^k v = v_k, \quad j = 0, \ldots, 3, \quad k = 0, 1 \quad \text{on } S = \Gamma \times (0,T), \tag{1.3}$$

where $\Gamma \subset \partial\Omega$, $\Gamma \in C^1$.

Let Q_- be a domain in $\mathbb{R}^2$ such that $S = \partial Q_- \cap \partial Q$, while Q_- and Q do not overlap. For any reasonable definition of a solution to the Cauchy problem (1.1)–(1.3), uniqueness in $Q_0 \subset Q$ will follow from the following uniqueness of the continuation principle: if (w, v) solves (1.1), (1.2) in $Q_- \cup S \cup Q$ and is zero on Q_-, then it is zero on Q_0. Observe that we generally cannot expect uniqueness in the whole Q due to finite speed of propagation of the hyperbolic component w.

First we consider a simple but important case of (real) analytic b_j, $b_{j,k}$ and show that the Holmgren–John theorem implies uniqueness in any subdomain Q_0 of Q formed of its points that can be reached by noncharacteristic (with respect to the operator $-\gamma\partial_t^2 + \Delta$) deformations of $\Gamma \times (0,T)$.

Theorem 1.1. *Let Q_* be a domain in $\mathbb{R}^3$. Let the coefficients b_j, $b_{j,k}$ be real analytic in Q_*. Let S_* be a C^2-surface inside Q_* which is time-like with respect to the wave operator $-\gamma\partial_t^2 + \Delta$ and which divides Q_* into two subdomains Q_- and Q_+. Let (w, v) be a (distribution) solution to the system* (1.1), (1.2) *in Q_*, which is zero in Q_+.*

Then (w, v) is zero near S_* in Q_*.

According to trace theorems, the Cauchy data for solutions with less regularity are in corresponding Sobolev spaces of negative order. Definition of the space $H_{(k-j-1/2)}(S)$ and applicability of the trace theorem for normal derivatives ∂_ν^j of functions from $H_{(k)}(\Omega)$ require $\Gamma \in C^{|k-j|}$.

Corollary 1.2. *Let $\Gamma \in C^2$. A solution $(w,v) \in C([0,T]; H_2(\Omega) \times L^2(\Omega))$ to the Cauchy problem* (1.1)–(1.3) *is uniquely determined at any point $(x,t) \in Q$ such that $\gamma^{1/2}d < t < T - \gamma^{1/2}d$, where d is the* (*Euclidean*) *distance in Q from x to Γ.*

We will prove these results in section 2. Meanwhile we observe that in Theorem 1.1 we consider more general domains Q_* than cylindrical Q and more general surfaces S_* than S. This generality is needed for noncharacteristic deformations to obtain Corollary 1.2. A C^1-surface in $\mathbb{R}^3$ with the unit normal ν is called time-like with respect to the wave operator $-\gamma\partial_t^2 + \Delta$ if $\nu_1^2 + \nu_2^2 > \gamma\nu_0^2$ at any point of this surface. d is the distance in Q, i.e., $d = \inf|\gamma|$ over all smooth curves $\gamma \subset Q$ joining x and a point of Γ; $|\gamma|$ is the Euclidean length of γ.

When coefficients are not analytic, this argument does not work and we will obtain uniqueness in a smaller (than in Corollary 1.2) domain Q_0 by using Carleman estimates. We let Q_ϵ be $Q \cap \{-\theta^2(t-T/2)^2 + |x-a|^2 - \rho > \epsilon\}$, where ρ is a parameter to be chosen later. We will give uniqueness results in two cases: (1) $\Gamma = \partial\Omega$, and (2) $\Omega \subset \{-h < x_2 < 0, |x_1| < r\}$, $\Gamma = \partial\Omega \cap \{x_2 < 0\}$. We will assume

$$\begin{aligned} &\Omega \subset B(0;\theta T/2), \quad \rho = 0 \quad \text{in case (1)}, \\ &4h(h+2a_2) < \theta^2 T^2, \quad \rho = a_2^2 + r^2, \quad a = (0,a_2) \quad \text{in case (2)}. \end{aligned} \tag{1.4}$$

Let $Q_\epsilon = Q \cap \{-\theta^2(t-T/2)^2 + |x-a|^2 - \rho > \epsilon\}$. Geometry of Q_ϵ (including illustrating figures) and the conditions (1.4) are discussed in Isakov [10, section 3.4]. In particular, choosing $\theta < \gamma^{-\frac{1}{2}}$ close to $\gamma^{-\frac{1}{2}}$ and $T > \frac{\mathrm{diam}\Omega}{\sqrt{\gamma}}$, one obtains a sharp uniqueness domain in case (1). In case (2) the uniqueness domain Q_0 is not optimal: the maximal space domain $Q_0 \cap \{t = T/2\}$ is never $\Omega \times \{T/2\}$, but it contains the domain $(\Omega \cap \{x_2 < -\frac{r^2}{2a_2}\}) \times \{T/2\}$, and selecting large a_2 and large T according to (1.4) (for example, greater than $2\sqrt{\gamma a_2}$), we can approximate Ω with any precision by uniqueness domains.

Theorem 1.3. *Let us assume that* (1.4) *is satisfied and that*

$$\gamma\theta^2 < 1. \tag{1.5}$$

Let the coefficients b_j, $b_{j,k}$, $b_{j,\alpha} \in L^\infty(Q)$.

Then any solution (w,v) to (1.1)–(1.3) *with lower order derivatives $\partial_t\Delta w$, $\partial_j\Delta w$, $\partial_t^2\partial_j w$, $\partial^\alpha w$, $|\alpha| \le 2$, $\partial_j v (j = 1,2)$ in $L^2(Q)$ is unique in Q_0.*

In the case of time independent coefficients, one can reduce regularity assumptions of Theorem 1.3 by using mollifying with respect to t and standard elliptic theory.

Corollary 1.4. *In addition to the conditions of Theorem* 1.3 *let us assume that the coefficients of* (1.1), (1.2) *do not depend on t and that the partial derivatives of first order of the coefficients $b_{1,k}$, $b_{2,k}$, $b_{5,k}$, $b_{6,k}$ are in $L^\infty(\Omega)$.*

Then any solution $(w,v) \in C([0,T]; H_2(\Omega) \times L^2(\Omega))$ to the Cauchy problem (1.1)–(1.3) *is unique in Q_0.*

We will prove these results in sections 2–5, where the following notation and known results are used.

We let $\zeta = \xi + i\tau\nabla_{t,x}\phi$, $\tau > 0$, $\xi = (\xi_0, \xi_1, \xi_2) \in \mathbb{R}^3$, and

$$\phi = e^{\lambda\psi}, \quad \psi(x,t) = -\theta^2(t-T/2)^2 + |x-a|^2 - \rho.$$

We have the following simple equalities:

$$
\begin{gathered}
\partial_0\phi = -\lambda\theta^2(t-T/2)\phi, \quad \partial_j\phi = \lambda(x-a)_j\phi,\\
\partial_0^2\phi = (\lambda^2\theta^4(t-T/2)^2 - \lambda\theta^2)\phi,\\
\partial_0\partial_j\phi = -\lambda^2\theta^2(t-T/2)(x-a)_j\phi, \quad j=1,2,\\
\partial_1\partial_2\phi = \lambda^2(x-a)_1(x-a)_2\phi, \quad \partial_j^2\phi = (\lambda^2(x-a)_j^2+\lambda)\phi.
\end{gathered} \tag{1.6}
$$

A differential quadratic from $\mathcal{G}v\overline{v}$ is the sum

$$\sum g^{\alpha\beta}(x,t)D^\alpha v\overline{D^\beta v},$$

and its symbol is

$$\sum g^{\alpha\beta}(x,t)\zeta^\alpha\overline{\zeta^\beta}.$$

In the next sections for a partial differential operator $P(x,t;D)$ with the principal symbol $p(x,t;\zeta)$ (which will be differently defined in sections 3 and 4), we will make use of the differential quadratic forms

$$\mathcal{F}(x,t;D,\overline{D},\tau) = |P(x,t;D+i\tau\nabla_{x,t}\phi(x,t))v|^2 - |\overline{P}(x,t;D-i\tau\nabla_{x,t}\phi(x,t))v|^2$$

and $\mathcal{G}$, which are obtained from $\mathcal{F}$ by integrating by parts and have the principal symbol

$$
\begin{aligned}
\mathcal{G}_{pr}(x,t;\xi,\xi,\tau) = 2\tau\sum\partial_j\partial_k\phi\partial p/\partial\zeta_j\overline{\partial p/\partial\zeta_k} - 2\mathcal{T}\sum\partial p/\partial\zeta_k\overline{\partial_k p}\\
- 2\mathcal{T}\sum\overline{p}(\partial^2 p/\partial x_k\partial\zeta_k + i\tau\partial^2 p/\partial\zeta_j\partial\zeta_k\partial_j\partial_k\phi),
\end{aligned} \tag{1.7}
$$

where $\partial p/\partial\zeta_j$, $\partial_k p\dots$ are calculated at the point $(x,t;\xi+i\tau\nabla_{x,t}\phi(x,t))$. For the operators Δ and $\partial_t-\Delta$, the sums are over $j,k=1,2$, and for the wave operator they are over j, $k=0,1,2$. The form $\mathcal{G}$ has the following useful property:

$$\int_Q \mathcal{G}v\overline{v} = \int_Q \mathcal{F}v\overline{v} \le \int_Q |P(\,;D+i\tau\nabla\phi)v|^2 \tag{1.8}$$

for all functions $v \in C_0^\infty(Q)$. As one can see from formula (1.7) and the similar formula for the symbol of $\mathcal{G}$ (where the principal symbol of P must be replaced by the complete symbol of P and for the heat operator the sum is over j, $k=0,1,2$),

$$\left|\int_Q(\mathcal{G}_{pr}-\mathcal{G})v\overline{v}\right| \le C\lambda\sum\|(\tau\lambda\phi)^{1-|\alpha|}\partial^\alpha v\|_{(0)}^2(Q), \tag{1.9}$$

where the sum is over $|\alpha|\le 1$ for the wave operator P with the additional condition $\alpha_0=0$ for the Laplace and heat operators.

We will use the abbreviation $\sigma = \lambda\tau\phi$.

2. Proof of uniqueness in the analytic case.

The proof of Theorem 1.1 is a reduction of the system (1.1), (1.2) to a first order system to which one can apply the Holmgren–John theorem.

Proof of Theorem 1.1. It suffices to show that a solution (w,v) is zero near any point of S_* which we can assume to be the origin. S_* near the origin is the graph of a C^2-function on its tangent plane, and Q_+ is the subgraph of this function. Replacing this function by its second order Taylor polynomial and subtracting the square of the distance to the origin, we can assume that S_* is analytic near the origin. Now one can find an analytic substitution with nonzero Jacobian at the origin transforming this surface into the surface $\{y_2 = 0\}$.

Expressing Δv from (1.2) and substituting into (1.1), we obtain a new system. Since S_* is time-like with respect to the wave operator, the new system in the new variables $y_0, \dots, y_2$ will have the form

$$\partial_2^4 W = A_4 W + A_1 V, \quad \partial_2^2 V = A_2 V + A_3 W \quad \text{near the origin,} \tag{2.1}$$

where A_j are linear partial differential operators of order j with coefficients analytic near the origin; moreover, A_4 does not contain the partial differentiation ∂_2^4 and A_2 does not contain the partial differentiation ∂_2^2. The new vector functions $\boldsymbol{W}$, $\boldsymbol{V}$ whose components are all partial derivatives of the function W up to order 3 and all partial derivatives of V up to order 1 satisfy the first order system of the Cauchy–Kowalevsky type

$$\partial_2 \boldsymbol{W} = \boldsymbol{A}(y; \boldsymbol{W}, \partial_0 \boldsymbol{W}, \partial_1 \boldsymbol{W}, \boldsymbol{V}), \quad \partial_2 \boldsymbol{V} = \boldsymbol{B}(y; \boldsymbol{W}, \boldsymbol{V}, \partial_0 \boldsymbol{V}, \partial_1 \boldsymbol{V})$$

near the origin. Here $\mathbf{A}$, $\mathbf{B}$ are matrices analytically depending on y near the origin. Indeed, if $W_j = \partial_2^k \partial^\alpha w$, where $k + |\alpha| = 3$, $k < 3$, $\alpha_2 = 0$, then $\partial^\alpha = \partial^{\alpha(*)} \partial_l$ for some $l < 2$ and $\partial_2 W_j = \partial_l W_m$, where $W_m = \partial_2^{k+1} \partial^{\alpha(*)} w$. If $W_j = \alpha_2^3 w$, then $\partial_2 W_j$ can be expressed from the first equation (2.1). Similarly, one obtains the equations for $\mathbf{V}$.

Since $\mathbf{W}$, $\mathbf{V}$ are zero on one side of S_*, by the Holmgren–John theorem [6], [11] we conclude that they are zero near the origin.

The proof is complete. □

Proof of Corollary 1.2. From the definition of the distance d it follows that there is a finite collection of intervals $I_1, \dots, I_m$ in Ω such that the starting point x^l of I_l is inside I_{l-1}, the starting point of I_1 is on Γ, the terminal point of I_m is x, and their total length $d^* = d_1 + \cdots + d_m$ satisfies the inequality $\gamma^{1/2} d_* < t < T - \gamma^{1/2} d_*$. We will use the "triangle lemma" 3.4.6 in [10] to propagate subsequently along these intervals. As one can see from its proof in [10] for scalar hyperbolic equations this lemma is valid for our system (1.1), (1.2) because in Theorem 1.1 we use time-like surfaces for the scalar hyperbolic operator. In our situation, this lemma says that if Tr is the triangle in the (x_2, t)-plane with the vertices $(0,0)$, $(-R, T_0)$, $(-R, T_0)$, and Tr_ϵ is the ϵ-perturbation $\{|x_1| < \epsilon\} \times Tr$ of Tr with respect to x_1 and $\gamma^{1/2} R < T_0$, then a solution to the Cauchy problem (1.1)–(1.3) with $\Gamma = \partial(Tr_\epsilon) \cap \{x_2 = -R\}$ is uniquely determined in the whole Tr_ϵ. Here ϵ is any positive number, so the triangle can be arbitrarily "thin."

We can choose ϵ so small that any ϵ-perturbation P_l of the rectangle $I_l \times (0, T)$ in the normal direction is still in Q. Using the triangle lemma for P_1 (with the choice of the x_2-axis parallel to I_1 and possible t-translations of triangles), we conclude that (w, v) is uniquely determined near the interval $\{(x^2, t) : \gamma^{1/2} d_1 < t < T - \gamma^{1/2} d_1\}$. Propagating along P_2 we conclude that (w, v) is unique near $\{(x^3, t) : \gamma^{1/2}(d_1 + d_2) < t < T - \gamma^{1/2}(d_1 + d_2)\}$. Repeating this step m times, we complete the proof. □

3. Carleman-type estimates with additional parameter for Laplace, wave, and plate operators.

Lemma 3.1. *There is C such that*

$$\lambda^{1/2} \| \sigma^{3/2 - |\alpha|} e^{\tau\phi} \partial^\alpha u_1 \|_{(0)}(Q) \le C \| e^{\tau\phi} \Delta u_1 \|_{(0)}(Q) \tag{3.1}$$

for all $u_1 \in C_0^\infty(Q_\epsilon)$ *provided* $|\alpha| \le 1$, $\alpha_0 = 0$, *and* $\tau > C(\lambda)$.

Proof. First we will derive Carleman estimates in $Q(s) = Q \cap \{t = s\}$ and then integrate them with respect to s over $(0, T)$.

We consider $p(\zeta) = \zeta \cdot \zeta$ and the corresponding form $\mathcal{G}$. Let $(x^0, t^0) \in \overline{Q_\epsilon}$. We will denote by ϕ^0, $\mathcal{G}^0, \ldots$ functions and forms at the point (x^0, t^0).

For brevity we will drop the index 0 until (3.5). We will prove that

$$|\zeta|^4 \le C(\lambda^{-1}\sigma\mathcal{G}_{pr} + C|p(\zeta)|^2). \tag{3.2}$$

Indeed, the equality $p(\zeta) = 0$ and the relations (1.6) imply that

$$|\xi|^2 = \tau^2|\nabla\phi|^2 = \sigma^2|x-a|^2, \quad \xi \cdot \nabla\phi = 0. \tag{3.3}$$

According to (1.6), (1.7)

$$\begin{aligned}\mathcal{G}_{pr}(\xi,\xi,\tau) &= 8\tau \sum \lambda^2\phi(x-a)_j(x-a)_k\zeta_j\overline{\zeta_k} + 8\lambda\phi\sum|\zeta_k|^2 \\ &\ge 8\tau\lambda^2\phi|(x-a)\cdot\zeta|^2 \ge 8\tau\lambda^4\phi^3\tau^2|x-a|^2 \ge \epsilon_1\lambda\sigma^3,\end{aligned}$$

where we kept only $\mathcal{T}\zeta$, and used (1.6) and the inequality $|x-a| > 0$ on $\overline{Q_\epsilon}$ due to the choice of the parameters s, a and to the definition of Q_ϵ. From (3.3) it follows that $|\zeta| \le C\sigma$, so we have

$$C\tau\phi\mathcal{G}_{pr}(\xi,\xi,\tau) \ge |\zeta|^4. \tag{3.4}$$

For continuity and homogeneity reasons, this inequality holds when $|p(\zeta)| \le \delta|\zeta|^2$ for some $\delta > 0$ not depending on λ, τ.

Indeed, using the notation $\tau^* = \lambda\tau(x-a)\phi$ from (1.7) as in the above computation we have

$$\begin{aligned}\tau\phi\mathcal{G}_{pr}(\xi,\xi,\tau) &\ge 8\sum\tau_j^*\tau_k^*\zeta_j\overline{\zeta_k} - 4\tau^2\phi\mathcal{T}(\overline{p}i\Delta\phi_0) \\ &= 8|\tau^*\cdot\zeta|^2 - 4\tau^2\lambda^2\phi^2\mathcal{T}(\overline{p}i(|x-a|^2 + 2/\lambda)) \ge 8|\tau^*|^4 - C|p|\,|\tau^*|^2,\end{aligned}$$

where we have used the formulae (1.6) and dropped the real part of $\tau^* \cdot \zeta$. The inequality $|p| \le \delta|\zeta|^2$ implies that $|\xi|^2 - |\tau^*|^2 \le \delta(|\xi|^2 + |\tau^*|^2)$, or $|\xi|^2 \le (1+\delta)/(1-\delta)|\tau^*|^2$, and hence $|\zeta|^2 \le 2/(1-\delta)|\tau^*|^2$. Choosing δ small and summing up, we obtain (3.4).

Consider the case $\delta|\zeta|^2 \le |p(\zeta)|$. The definition of $\mathcal{G}$ implies that in any event $\tau\phi\mathcal{G} \ge -C_1|\zeta|^4$. Using this inequality and choosing $C > 2C_1/\delta$, we complete the proof of (3.2).

From (3.2) and from the property (1.9), we conclude that

$$\begin{aligned}\sum\int_{Q(s)}\sigma_0^{4-2|\alpha|}|\partial^\alpha v|^2 \le C\Bigg(&\lambda^{-1}\sigma_0\int_{Q(s)}\mathcal{G}(x^0,t^0;D,\overline{D},\tau)v\overline{v} \\ &+ \int_{Q(s)}|P(D+i\tau\nabla\phi^0)v|^2\Bigg)\end{aligned} \tag{3.5}$$

provided $v \in C_0^\infty(Q_\epsilon)$, $|\alpha| \le 2$, $\alpha_0 = 0$, and $\tau > C$. From the definition of $\mathcal{G}$ and ϕ it follows that

(3.6)

$$\left|\int_{Q(s)}(\mathcal{G}(x,t;D,D,\tau) - \mathcal{G}(x^0,t^0;D,\overline{D},\tau))v\overline{v}\right| \le \omega(\delta,\lambda)\sum(\tau\phi)^{3-2|\alpha|}\int_{Q(s)}|\partial^\alpha v|^2,$$

where the sum is over $|\alpha| \leq 1$, $\alpha_0 = 0$,

(3.7)
$$\left| \int_{Q(s)} |P(D + i\tau\nabla\phi^0)v|^2 - |P(D + i\tau\nabla\phi)|^2 \right| \leq \omega(\delta, \lambda) \sum (\tau\phi^0)^{4-2|\alpha|} \int_{Q(s)} |\partial^\alpha v|^2,$$

where the sum is over $|\alpha| \leq 2$, $\alpha_0 = 0$, and $\omega(\delta, \lambda) \to 0$ as $\delta \to 0$ when λ is fixed, provided $v \in C_0^\infty(Q_\epsilon \cap B(x^0, t^0, \delta))$. Given $\lambda > 1$ we can use (3.6), (3.7) and choose $\delta > 0$ so small that differences between right sides of (3.5) with (x, t) and (x^0, t^0) are absorbed by the left side. This yields

$$\begin{aligned}
&\sum \int_{Q(s)} \sigma_0^{4-2|\alpha|} |\partial^\alpha v|^2 \\
&\quad \leq \; C\left(\lambda^{-1}\sigma_0 \int_{Q(s)} \mathcal{G}(x, t; D, \overline{D}, \tau)v\overline{v} + \int_{Q(x)} |P(D + i\tau\nabla\phi)v|^2 \right) \\
&\quad \leq \; C\lambda^{-1}\sigma_0 \int_{Q(s)} |P(D + i\tau\nabla\phi)v|^2 \qquad (3.8)
\end{aligned}$$

due to the property (1.8) of the form $\mathcal{G}$. Dividing (3.8) by σ_0 and using as above that $\sigma_0 = \sigma(1 + \omega(\delta, \lambda))$, we can replace σ_0 in the left side by σ. Returning to $u_1 = e^{\tau\phi}v$, we obtain the inequality

$$\lambda \sum \int_{Q(s)} \sigma^{3-2|\alpha|} e^{2\tau\phi} |\partial^\alpha u_1|^2 \leq C \int_{Q(s)} e^{2\tau\phi} |\Delta u_1|^2$$

for all $u_1 \in C_0^\infty(B(x^0, t^0; \delta))$ for some δ and $\lambda > C$, $\tau > C(\lambda)$. Here the sums are over $|\alpha| \leq 2$, $\alpha_0 = 0$. Integrating with respect to s over $(0, T)$, we obtain (3.1) for all u_1 with small support. Using the partition of the unity over $\overline{Q_\epsilon}$, we complete the proof. □

Lemma 3.2. *Assume that*

$$\gamma\theta^2 < 1. \qquad (3.9)$$

Then there is C such that

$$\|\sigma^{3/2-|\alpha|} e^{\tau\phi} \partial^\alpha u_2\|_{(0)}(Q) \leq C \|e^{\tau\phi}(-\gamma\partial_t^2 + \Delta)u_2\|_{(0)}(Q) \qquad (3.10)$$

for all functions $u_2 \in C_0^\infty(Q_\epsilon)$, $\tau > C(\lambda)$, $|\alpha| \leq 1$.

Proof. We will adjust the method of the proof of Lemma 3.1 to the wave operator with the symbol $p(\widetilde{\zeta}) = \gamma\zeta_0^2 - \zeta \cdot \zeta$. We let $\zeta_0 = \xi_0 + i\tau\partial_t\phi$, $\widetilde{\zeta} = (\zeta_0, \zeta)$, $X = (x, t)$ and will use most of the notation of Lemma 3.1. In particular, we will fix any $X^0 \in \overline{Q_\epsilon}$.

We claim that

$$|\zeta_0|^2 + |\zeta|^2 \leq C(\sigma_0^{-1} \mathcal{G}^0_{pr}(\widetilde{\xi}, \widetilde{\xi}, \tau) + C\lambda^2 |p(\widetilde{\zeta})|^2 |\widetilde{\zeta}|^{-2}). \qquad (3.11)$$

In the proof of (3.11), we will drop the index 0.

Let $p(\widetilde{\zeta}) = 0$. Using (1.6) we will have

$$\gamma\xi_0^2 - |\xi|^2 = \sigma(|x - a|^2 - \theta^4(t - T/2)^2), \qquad (3.12)$$

and the formulas (1.6), (1.7) give

$$\begin{aligned}\mathcal{G}_{pr}(\widetilde{\xi},\widetilde{\xi},\tau) &= 2\tau(\partial_0^2\phi 4|\zeta_0|^2 - 4\sum \partial_0\partial_j\phi(\zeta_0\overline{\zeta_j} + \overline{\zeta_0}\zeta_j) + 4\sum \partial_j\partial_k\phi\zeta_j\overline{\zeta_k})\\ &= 8\sigma(-\gamma^2\theta^2|\zeta_0|^2 + |\zeta|^2 + \lambda|\gamma\theta^2(t-T/2)\zeta_0 + (x-a)\cdot\zeta|^2).\end{aligned}$$

To obtain (3.11) we will again use homogeneity and continuity arguments assuming $|\widetilde{\zeta}|^2 = 1$.

When $\tau = 0$, we have $\xi_0^2 + |\xi|^2 = 1$, $\gamma\xi_0^2 = |\xi|^2$. So

$$-\gamma^2\theta^2\xi_0^2 + |\xi|^2 = \gamma(1-\gamma\theta^2)/(1+\gamma) > C^{-1}$$

according to assumption (3.9). Hence in this case

$$\sigma^{-1}\mathcal{G}_{pr}(\widetilde{\xi},\widetilde{\xi},\tau) \geq C^{-1}|\widetilde{\zeta}|^2.$$

As in Lemma 3.1, we generally have

$$\sigma^{-1}\mathcal{G}_{pr} \geq 8(-\gamma^2\theta^2|\zeta_0|^2 + |\zeta|^2 + \lambda|\gamma\theta^2(t-T/2)\zeta_0 + (x-a)\cdot\zeta|^2) - C\lambda|p(\widetilde{\zeta})|. \tag{3.13}$$

Let

$$|p(\widetilde{\zeta})| < \delta\lambda^{-1}|\widetilde{\zeta}|^2. \tag{3.14}$$

Then we have $-\delta\lambda^{-1}|\widetilde{\zeta}|^2 < \gamma\xi_0^2 - |\xi|^2 - (\gamma\tau_0^{*2} - |\tau^*|^2) < \delta\lambda^{-1}|\widetilde{\zeta}|^2$.

Consider the case $|\widetilde{\tau}^*|^2 < \delta|\widetilde{\zeta}|^2$. Using the above inequality we obtain $\gamma\xi_0^2 - ((\gamma+1)\delta + \delta\lambda^{-1})|\widetilde{\zeta}|^2 < |\xi|^2$. Using the last inequality, (3.9) and (3.14), we will have

$$\begin{aligned}-\gamma^2\theta^2|\zeta_0|^2 + |\zeta|^2 &\geq \gamma(1-\gamma\theta^2)\xi_0^2 - ((2\gamma+1)\delta + \delta\lambda^{-1})|\widetilde{\zeta}|^2\\ &\geq C^{-1}\xi_0^2 - (C\delta + \delta\lambda^{-1})|\widetilde{\zeta}|^2 > C^{-1}|\widetilde{\zeta}|^2 - (C\delta + \delta_0\lambda^{-1})|\widetilde{\zeta}|^2.\end{aligned}$$

So for some small δ, (3.13) and (3.14) imply (3.11). From now on we will fix such δ.

When $\delta|\widetilde{\zeta}|^2 \leq |\widetilde{\tau}^*|^2$, then $|\widetilde{\zeta}|^2 \leq C\sigma^2$. Using that $|\zeta| \geq |\mathcal{T}\zeta|$ and again using (1.6), we obtain

$$\begin{aligned}|\gamma\theta^2(t-T/2)\zeta_0 + (x-a)\cdot\zeta|^2 &\geq |-\gamma\theta^4\lambda(t-T/2)^2\tau\phi + \tau\lambda|x-a|^2\phi^2|^2\\ &= \sigma^2|-\gamma\theta^4(t-T/2)^2 + |x-a|^2| > \epsilon\sigma^2\end{aligned}$$

due to the condition (3.9) and to the definition of Q_ϵ. Hence from (3.13) and (3.14), we have

$$\sigma^{-1}\mathcal{G}^0_{pr} \geq -C|\widetilde{\zeta}|^2 + \lambda C^{-1}|\widetilde{\zeta}|^2 - C\delta|\widetilde{\zeta}|^2,$$

which implies (3.11) when $\lambda > C$.

Now we will consider the remaining case $|p(\widetilde{\zeta})| \geq \delta\lambda^{-1}|\widetilde{\zeta}|^2$. Using (3.13), we conclude that

$$\begin{aligned}\sigma^{-1}\mathcal{G}^0_{pr} + C_1\lambda^2|p(\widetilde{\zeta})|^2|\widetilde{\zeta}|^{-2} &\geq -C|\widetilde{\zeta}|^2 + \lambda|p(\widetilde{\zeta})|(C_1\lambda|p(\widetilde{\zeta})|\,|\widetilde{\zeta}|^{-2} - C)\\ &\geq -C|\widetilde{\zeta}|^2 + \delta|\widetilde{\zeta}|^2(C_1\delta - C),\end{aligned}$$

and choosing C_1 sufficiently large we complete the proof of (3.11).

The inequality (3.11) for the principal symbol of the differential quadratic form and the bound (1.9) as in the proof of Lemma 3.1 imply that

(3.15)

$$\sum_{|\alpha|\le 1}\int_Q \sigma_0^{2-2|\alpha|}|\partial^\alpha v|^2 \le C\left(\sigma_0^{-1}\int_Q \mathcal{G}^0_{pr}(D,\overline{D},\tau)v\overline{v} + \lambda^2\int |p(\widetilde{\zeta})|^2|\widetilde{\zeta}^0|^{-2}|\widehat{v}(\widetilde{\xi})|^2\,d\widetilde{\xi}\right).$$

As in Lemma 3.1,

$$\left|\int_Q (\mathcal{G}-\mathcal{G}^0)v\overline{v}\right| \le \omega(\delta,\lambda)\sum_{|\alpha|\le 1}\tau^{3-2|\alpha|}\int_Q |\partial^\alpha v|^2.$$

Letting

$$|||v|||_k = \left(\int |\widetilde{\zeta}^0|^{2k}|\widehat{v}(\widetilde{\xi})|^2\,d\widetilde{\xi}\right)^{1/2}$$

and using that by Lemma 8.4.1 in [5] and Lemma 2.1 in [7]

$$\begin{aligned}&|||P(D+i\tau(\nabla_{x,t}\phi)^0)v|||^2_{-1}\\ &\quad\le 2|||P(D+i\tau\nabla_{x,t}\phi)v|||^2_{-1} + \omega(\delta,\tau;\lambda)\sum_{|\alpha|\le 1}\int_Q \sigma_0^{2-2|\alpha|}|\partial^\alpha v|^2\end{aligned}$$

when $v \in C_0^\infty(B(X^0;\delta))$, where $\omega \to 0$ as $\delta \to 0$, $\tau \to 0$, and λ is fixed. Choosing δ small, absorbing the differences at X and X^0 as in the proof of Lemma 3.1, and observing that

$$\lambda^2|||P(D+i\tau\nabla_{x,t}\phi)v|||^2_{-1} \le C\tau^{-2}\|P(D+i\tau\nabla_{x,t}\phi)v\|^2_{(0)},$$

we derive from (3.15)

$$\begin{aligned}\sum_{|\alpha|\le 1}\int_Q \sigma_0^{2-2|\alpha|}|\partial^\alpha v|^2 &\le C\left(\sigma_0^{-1}\int_Q \mathcal{G}v\overline{v} + C\tau^{-2}\int_Q |P(D+i\tau\nabla_{x,t}\phi)v|^2\right)\\ &\le C(\sigma_0^{-1}+\tau^{-2})\int_Q |P(D+i\tau\nabla_{x,t}\phi)v|^2\end{aligned}$$

due to the property (1.8) of the form $\mathcal{G}$. Choosing $\tau > \lambda$ large we can achieve that $\sigma_0^{-1} > \tau^{-2}$, and therefore we can drop the term with τ in front of the last integral. Arguing as at the end of the proof of Lemma 3.1, we can replace σ_0 by σ, return to the function u_2, and use partition of the unity to complete the proof. □

A "substitution" of the estimate of Lemma 3.1 into the estimate of Lemma 3.2 gives the following lemma.

Lemma 3.3. *Under the condition* (3.9) *there is* C *such that*

(3.16)

$$\begin{aligned}&\|\sigma^{3-|\beta|}e^{\tau\phi}\partial^\beta w_0\|_{(0)}(Q) + \|\sigma^{1/2}e^{\tau\phi}\partial_t^2\partial^\alpha w_0\|_{(0)}(Q) + \|\sigma^{1/2}e^{\tau\phi}\partial_t\Delta w_0\|_{(0)}(Q)\\ &\qquad + \|\sigma^{1/2}e^{\tau\phi}\partial^\alpha\Delta w_0\|_{(0)}(Q) \le C\|e^{\tau\phi}(-\gamma\Delta\partial_t^2+\Delta^2)w_0\|_{(0)}(Q)\end{aligned}$$

for all $w_0 \in C_0^\infty(Q_\epsilon)$ *provided* $|\beta| \le 2$, $|\alpha| = 1$, $\alpha_0 = 0$, *and* $\tau > C(\lambda)$.

Proof. This lemma follows from Lemmas 3.1 and 3.2 by letting $u_1 = (-\gamma\partial_t^2 + \Delta)w_0$, $u_2 = \Delta w_0$, and choosing large τ to eliminate terms resulting from commuting multiplication by ϕ and the differential operators.

Indeed, letting $u_1 = (-\gamma\partial_t^2 + \Delta)w_0$ in (3.1), we obtain

(3.17)
$$\begin{aligned}&\|e^{\tau\phi}(-\gamma\partial_t^2 + \Delta)\Delta w_0\|_{(0)}(Q)\\ &\quad\ge\ \lambda/C \sum_{|\alpha^*|\le 1, \alpha_0^*=0} \|\sigma^{3/2-|\alpha^*|} e^{\tau\phi}\partial^{\alpha^*}(-\gamma\partial_t^2 + \Delta)w_0\|_{(0)}(Q).\end{aligned}$$

By Lemma 3.2 with $u_2 = \sigma^{3/2}w_0$ we have

$$(3.18)\ \sum_{|\alpha^{**}|\le 1} \|\sigma^{3/2-|\alpha^{**}|}e^{\tau\phi}\partial^{\alpha^{**}}(\sigma^{3/2}w_0)\|_{(0)} \le C\|e^{\tau\phi}(-\gamma\partial_t^2 + \Delta)(\sigma^{3/2}w_0)\|_{(0)}.$$

By the Leibniz formula

$$\begin{aligned}\partial_j(\sigma^{3/2}w_0) &= \tau^{3/2}b_j w_0 + \sigma^{3/2}\partial_j w_0,\\ (-\gamma\partial_t^2 + \Delta)(\sigma^{3/2}w_0) &= \sigma^{3/2}(-\gamma\partial_t^2 + \Delta)w_0 + \tau^{3/2}\sum_{|\beta|\le 1} b_\beta \partial^\beta w_0,\end{aligned}$$

where b are bounded functions determined only by ψ and λ. Using the triangle inequality, we obtain from (3.18) that

$$\begin{aligned}&\sum_{|\alpha^{**}|\le 1} \|\sigma^{3-|\alpha^{**}|}e^{\tau\phi}\partial^{\alpha^{**}}w_0\|_{(0)} - \tau^2 C(\lambda)\|e^{\tau\phi}w_0\|_{(0)}\\ &\quad\le\ C\|\sigma^{3/2}e^{\tau\phi}(-\gamma\partial_t^2 + \Delta)w_0\|_{(0)} + C(\lambda)\tau^{3/2}\sum_{|\beta|\le 1}\|e^{\tau\phi}\partial^\beta w_0\|_{(0)}.\end{aligned}$$

Choosing large τ will absorb the second terms of the left and right sides by the first term in the left side.

Similarly we have

$$\sum \|\sigma^{2-|\alpha^{**}|}e^{\tau\phi}\partial^{\alpha^*+\alpha^{**}}w_0\|_{(0)} \le C\sum \|\sigma^{1/2}e^{\tau\phi}(-\gamma\partial_t^2\partial^{\alpha^*} + \Delta\partial^{\alpha^*})w_0\|_{(0)},$$

where the sums are over $|\alpha^*| = 1$, $\alpha_0^* = 0$, $|\alpha^{**}| \le 1$, $\tau > C(\lambda)$.

The last two inequalities combined with (3.17) yield

$$\sum_{|\beta|\le 2} \|\sigma^{3-|\beta|}e^{\tau\phi}\partial^\beta w_0\|_{(0)} \le C/\lambda\|e^{\tau\phi}(-\gamma\partial_t^2 + \Delta)\Delta w_0\|_{(0)}.$$

Indeed, when $|\beta_0| \le 1$, it follows directly from these inequalities by letting $\beta = \alpha^* + \alpha^{**}$. When $\beta_0 = 2$, the partial derivative $\partial^\beta w_0 = 1/\gamma(\gamma\partial_t^2 - \Delta)w_0 + 1/\gamma\Delta w_0$ can be bounded from the right side of (3.17) ($\alpha^* = 0$) and the already obtained bounds on space derivatives of second order.

To complete the proof we similarly use Lemma 3.2 with $u_2 = \Delta w_0$ bounding the last two terms of the left side in (3.16). Expressing as above $\partial_t^2 w_0$ as the sum of the wave operator and the Laplacian and utilizing again the right side of (3.17) with $|\alpha^*| = 1$ and previous bounds of Δw_0, we obtain (3.16). □

4. Carleman estimates for the heat equation. In this section we consider $p(\widetilde{\zeta}) = i\zeta_0 + \zeta \cdot \zeta$ and will accordingly modify the proof of Lemma 3.1.

LEMMA 4.1. *There is C such that*

$$\lambda^{1/2}\|\sigma^{3/2-|\alpha|}e^{\tau\phi}\partial^\alpha u\|_{(0)}(Q) \le C\|e^{\tau\phi}(\partial_t - \Delta)u\|_{(0)}(Q) \tag{4.1}$$

for all $u \in C_0^\infty(Q_\epsilon)$, $|\alpha| \le 2$, $\alpha_0 = 0$, provided $\tau > C(\lambda)$.

Proof. As in Lemma 3.1, we first fix $X^0 \in \overline{Q_\epsilon}$ and obtain local estimates on functions supported near this point. As above, for a while we drop the index 0.

We claim that for some C

$$|\zeta_0|^2 + |\zeta|^4 \le C(\lambda^{-1}\sigma\mathcal{G}_{pr} + C|p(\widetilde{\zeta})|^2). \tag{4.2}$$

Indeed, the equality $p(\widetilde{\zeta}) = 0$ and the relations (1.6) imply that

$$\sigma\theta^2(t - T/2) + |\xi|^2 = \sigma^2|x-a|^2, \quad \xi_0 = 2\sigma\xi \cdot (x-a) \tag{4.3}$$

and

$$\begin{aligned}\mathcal{G}_{pr} &= 8\sigma\lambda\sum(x-a)_j(x-a)_j\zeta_j\overline{\zeta_k} + 8\lambda\phi\sum|\zeta_k|^2\\ &\ge 8\lambda\sigma|(x-a)\cdot\zeta|^2 \ge 8\lambda\sigma^3|x-a|^2 \ge 8C^{-1}\lambda\sigma^3,\end{aligned} \tag{4.4}$$

provided $X \in Q_\epsilon$. We can assume that $\sigma > C$; then the first equality of (4.3) implies that $|\xi| \le C\sigma$, and using in addition the second equality of (4.3), we conclude that $|\xi_0| \le C\sigma^2$, so $|\zeta_0| \le C\sigma^2$. Now, from (4.4) it follows that

$$\lambda^{-1}\sigma\mathcal{G}_{pr} \ge C^{-1}(|\zeta_0|^2 + |\zeta|^4). \tag{4.5}$$

Modifying the homogeneity and continuity arguments from the proof of Lemma 3.1, we conclude that (4.5) remains valid when $|p(\widetilde{\zeta})|^2 \le \delta(|\zeta_0|^2 + |\zeta|^4)$ for some small positive δ.

When $\delta(|\zeta_0|^2 + |\zeta|^4) \le |p(\widetilde{\zeta})|^2$, we can with no changes repeat the argument in the elliptic case and complete the proof of (4.2).

From (4.2) and from the property (1.9) of differential quadratic forms, we obtain

$$\sigma_0^{4-2|\alpha|}\int_Q |\partial^\alpha v|^2 + \int_Q |\partial_0 v|^2 \le C\left(\lambda^{-1}\sigma_0\int_Q \mathcal{G}^0 v\overline{v} + \int_Q |P(D + \tau i\nabla_{x,t}\phi^0)v|^2\right) \tag{4.6}$$

provided $\tau > C$, $|\alpha| \le 2$, $\alpha_0 = 0$, $v \in C_0^\infty(Q_\epsilon)$. As above, from the definition of $\mathcal{G}$ and from the regularity assumptions on ϕ, it follows that

$$\left|\int_Q (\mathcal{G} - \mathcal{G}^0)v\overline{v}\right| \le \omega(\delta;\lambda)\left(\sum_{|\alpha|\le 1}\tau^{3-2|\alpha|}\int_Q |\partial^\alpha v|^2\right),$$

$$\left|\int_Q\right| P(D + i\tau\nabla_{x,t}\phi)v|^2 - |P(D + i\tau\nabla_{x,t}\phi^0)v|^2| \le \omega(\delta;\lambda)\sum_{|\alpha|\le 2,\alpha_0=0}\tau^{4-2|\alpha|}\int_Q |\partial^\alpha v|^2$$

provided $v \in C_0^\infty(B(X^0;\delta))$, where $\omega(\delta;\lambda) \to 0$ as $\delta \to 0$ and λ is fixed. Using these inequalities, we can replace $\mathcal{G}^0$ and ϕ^0 in (4.6) by $\mathcal{G}$ and ϕ and remove $\mathcal{G}$ as in the elliptic case. Observe that in the parabolic case we handle all α with $|\alpha| \le 2$, $\alpha_0 = 0$.

Similarly, returning to $v_0 = e^{\tau\phi}v$ and using partition of the unity we complete the proof of Lemma 4.1. $\square$

Corollary 4.2. *Under the conditions of Lemma* 4.1*, we have*

$$\lambda^{1/2}\|\sigma^{2-|\alpha|}e^{\tau\phi}\partial^{\alpha}v_0\|_{(0)}(Q) \leq C\|\sigma^{1/2}e^{\tau\phi}(\partial_t - \Delta)v_0\|_{(0)}(Q) \tag{4.7}$$

for all $v_0 \in C_0^{\infty}(Q_{\epsilon})$*,* $|\alpha| \leq 2$*,* $\alpha_0 = 0$*, provided* $\tau > C(\lambda)$*.*

Proof. As in the proof of Lemma 3.3, letting in (4.1) $v_0 = \sigma^{1/2}v^*$ and using the Leibniz formula and the triangle inequality, we will have

$$\lambda^{1/2}\sum_{|\alpha|\leq 2}\|\sigma^{2-|\alpha|}e^{\tau\phi}\partial^{\alpha}v^*\|_{(0)} - C(\lambda)\sum_{|\alpha|\leq 1}\tau^{1-|\alpha|}\|e^{\tau\phi}\partial^{\alpha}v^*\|_{(0)}$$
$$\leq C\left(\|\sigma^{1/2}e^{\tau\phi}(\partial_t - \Delta)v^*\|_{(0)} + C(\lambda)\sum\tau^{1/2}\|e^{\tau\phi}\partial^{\alpha}v^*\|_{(0)}\right).$$

Choosing $\tau > C(\lambda)$, we absorb the second sums in the left and right sides by the first sum in the left side and complete the proof of Corollary 4.2. $\square$

5. Proofs of main results in the nonanalytic case. The Carleman-type estimates like (3.1), (3.10), (3.16), (4.1), and (4.7) are classical (and the only available) tools to prove uniqueness in the Cauchy problem for corresponding partial differential equations. Their introduction and use comes back to the pioneering work of T. Carleman of 1938. We refer to discussions of this method in Hörmander [5, sections 8.1–8.3] and Isakov [10, section 3.2]. Below we will use the Carleman method. The introduction of large parameter τ in the weight function $e^{\tau\phi}$ with ϕ decaying away from S is the crucial idea of this method. It helps to neglect parts of the boundary where no data are available. If we use Lemma 3.3 and Corollary 4.2 without the additional large parameter λ (which were known), we arrive at the inequalities (5.2), (5.3), but when bounding Δv_0 in the right side of (5.2) from (5.3) ($|\alpha| = 2$), we will lose the large parameter and will not be able to eliminate the terms with $\partial_t\Delta w$ in the right side and therefore complete the proof.

Proof of Theorem 1.3. We will introduce a cut-off function $\chi \in C_0^{\infty}(\mathbb{R}^3)$ such that it is 1 on $Q_{2\epsilon}$ and 0 on $Q\backslash Q_{\epsilon}$. Let $w_0 = \chi w$, $v_0 = \chi v$. By using the Leibniz formula, we derive from (1.1), (1.2) that

(5.1)

$$-\gamma\Delta\partial_t^2 w_0 + \Delta^2 w_0$$
$$= -\Delta v_0 + L_1(x,t;w,\partial_t w,\nabla w,\nabla^2 w,\nabla\partial_t w,\partial_t^2 w,\nabla\partial_t^2 w,\partial_t\Delta w,\nabla\Delta w,v,\nabla v),$$
$$\partial_t v_0 - b\Delta v_0 = L_2(x,t;w,\nabla w,\partial_t w,\nabla\partial_t w,\Delta w,\nabla\Delta w_0,\partial_t\Delta w_0,\nabla\partial_t^2 w_0,v,\nabla v),$$

where L_1, L_2 are linear functions of $w,\ldots,\nabla v$ with the coefficients in $L^{\infty}(Q)$.

Applying Lemma 3.3 and expressing the left side from the first equation through its right side, we will have the Carleman-type estimates

(5.2)

$$\sum(\|\sigma^{3-|\beta|}e^{\tau\phi}\partial^{\beta}w_0\|^2_{(0)}(Q_{\epsilon}) + \|\sigma^{1/2}e^{\tau\phi}\nabla\partial_t^2 w_0\|^2_{(0)}(Q_{\epsilon}) + \|\sigma^{1/2}e^{\tau\phi}\nabla_{x,t}\Delta w_0\|^2_{(0)}(Q_{\epsilon}))$$
$$\leq\ C(\|e^{\tau\phi}\Delta v_0\|^2_{(0)}(Q_{\epsilon}) + \sum(\|e^{\tau\phi}\partial^{\beta}w\|^2_{(0)}(Q_{\epsilon}) + \|e^{\tau\phi}\nabla_{x,t}\Delta w\|^2_{(0)}(Q_{\epsilon})$$
$$+\ \|e^{\tau\phi}\nabla\partial_t^2 w\|^2_{(0)}(Q_{\epsilon}) + \|e^{\tau\phi}\partial^{\alpha}v\|^2_{(0)}(Q_{\epsilon})))$$

when $\tau > C(\lambda)$, where the sums are over $|\beta| \leq 2$, $|\alpha| = 1$, $\alpha_0 = 0$. Similarly, from Corollary 4.2

$$\begin{aligned}\lambda \sum \|\sigma^{2-|\beta^*|} e^{\tau\phi} \partial^{\beta^*} v_0\|^2_{(0)}(Q_\epsilon) &\leq C \sum (\|\sigma^{1/2} e^{\tau\phi} \nabla_{x,t} \Delta w\|^2_{(0)}(Q_\epsilon) \\ &+ \|\sigma^{1/2} e^{\tau\phi} \nabla \partial_t^2 w\|^2_{(0)}(Q_\epsilon) + \|\sigma^{1/2} e^{\tau\phi} \partial^\beta w\|^2_{(0)}(Q_\epsilon) + \|\sigma^{1/2} e^{\tau\phi} \partial^\alpha v\|^2_{(0)}(Q_\epsilon))\end{aligned} \tag{5.3}$$

for the same τ, α, β, and $|\beta^*| \leq 2$, $b_0^* = 0$.

We will add (5.2) and (5.3) multiplied by $\lambda^{-1/2}$ to obtain

$$\begin{aligned}&\sum (\|\sigma^{3-|\beta|} e^{\tau\phi} \partial^\beta w_0\|^2_{(0)}(Q_\epsilon) + \|\sigma^{\frac{1}{2}} e^{\tau\phi} \nabla \partial_t^2 w_0\|^2_{(0)}(Q_\epsilon) + \|\sigma^{\frac{1}{2}} e^{\tau\phi} \nabla_{x,t} \Delta w_0\|^2_{(0)}(Q_\epsilon) \\ &\quad + \lambda^{\frac{1}{2}} \|\sigma^{2-|\beta^*|} e^{\tau\phi} \partial^{\beta^*} v_0\|^2_{(0)}(Q_\epsilon)) \leq C(\|e^{\tau\phi} \Delta v_0\|^2_{(0)}(Q_\epsilon) \\ &\quad + \sum (\|e^{\tau\phi} \partial^\beta w\|^2_{(0)}(Q_\epsilon) + \|e^{\tau\phi} \nabla_{x,t} \Delta w\|^2_{(0)}(Q_\epsilon) + \|e^{\tau\phi} \nabla \partial_t^2 w\|^2_{(0)}(Q_\epsilon) \\ &\quad + \|e^{\tau\phi} \partial^\alpha v\|^2_{(0)}(Q_\epsilon)) + C\lambda^{-\frac{1}{2}} \sum (\|\sigma^{\frac{1}{2}} e^{\tau\phi} \nabla_{x,t} \Delta w\|^2_{(0)}(Q_\epsilon) + \|\sigma^{\frac{1}{2}} e^{\tau\phi} \nabla \partial_t^2 w\|^2_{(0)}(Q_\epsilon) \\ &\quad + \|\sigma^{\frac{1}{2}} e^{\tau\phi} \partial^\beta w\|^2_{(0)}(Q_\epsilon) + \|\sigma^{\frac{1}{2}} e^{\tau\phi} \partial^\alpha v\|^2_{(0)}(Q_\epsilon)).\end{aligned}$$

We will break Q_ϵ into $Q_{2\epsilon}$ and its complement $Q_\epsilon \backslash Q_{2\epsilon}$ and choose sufficiently large λ to absorb the integral of Δv_0 in the right side by the last sum in the left side and to absorb the integrals of $\sigma^{\frac{1}{2}} \nabla_{x,t} \Delta w$, $\sigma^{\frac{1}{2}} \nabla \partial_t^2 w$ over $Q_{2\epsilon}$ (where $w_0 = w$, $v_0 = v$) in the right side by the corresponding integrals in the left side. Then we fix this λ, shrink the integration domain in the left side to $Q_{2\epsilon}$, and choose sufficiently large τ to absorb the integrals of the right side over $Q_{2\epsilon}$ by the integrals in the left side to obtain

$$\begin{aligned}&\tau(\|e^{\tau\phi} w\|^2_{(0)}(Q_{2\epsilon}) + \|e^{\tau\phi}\|^2_{(0)}(Q_{2\epsilon})) \leq C\tau^{1/2} \sum (\|e^{\tau\phi} \partial^\beta w\|^2_{(0)}(Q_\epsilon \backslash Q_{2\epsilon}) \\ &\quad + \|e^{\tau\phi} \nabla \Delta w\|^2_{(0)}(Q_\epsilon \backslash Q_{2\epsilon}) + \|e^{\tau\phi} \partial_t \Delta w\|^2_{(0)}(Q_\epsilon \backslash Q_{2\epsilon}) + \|e^{\tau\phi} \partial_t^2 \partial^\alpha w\|^2_{(0)}(Q_\epsilon \backslash Q_{2\epsilon}) \\ &\quad + \|e^{\tau\phi} \partial^\alpha v\|^2_{(0)}(Q_\epsilon \backslash Q_{2\epsilon})).\end{aligned}$$

Using that $\phi^* = \sup \phi$ over $Q_\epsilon \backslash Q_{2\epsilon}$ is equal to $\inf \phi$ over $Q_{2\epsilon}$, replacing ϕ by ϕ^* in both sides of the inequality, and dividing by $\tau^{1/2} e^{2\tau\phi^*}$, we arrive at

$$\tau^{1/2}(\|w\|^2_{(0)}(Q_{2\epsilon}) + \|v\|^2_{(0)}(Q_{2\epsilon})) \leq CM,$$

where M is the sum of $L^2(Q)$-norms of all partial derivatives of w and v entering the right side of the last inequality. Letting $\tau \to \infty$, we conclude that $w = v = 0$ on $Q_{2\epsilon}$ for any $\epsilon > 0$.

The proof is complete. □

Proof of Corollary 1.4. Let $\epsilon > 0$. The mollified functions $w_\delta = \chi_\delta * w$, $v_\delta * v$, where * denotes convolution with respect to t and χ_δ is the standard mollifying kernel [6, sections 1.2, 4.1–4.3], are well defined in $Q \cap \{\delta < t < T - \delta\}$ and solve there the Cauchy problem (1.1)–(1.3) with zero Cauchy data on S. From well-known properties of mollifiers and the assumptions on (w, v), we have $(\partial_t^k w_\delta, \partial_t^k v_\delta) \in C([0, T];$ $H_2(\Omega) \times L^2(\Omega))$. We can find small δ and a domain $\Omega^\bullet$ with $\overline{\Omega^\bullet} \subset \Omega \cup \Gamma$ so that $\overline{Q}_{\epsilon/2} \subset (\delta, T - \delta) \times \Omega^\bullet$. Transferring all terms of the equations (1.1), (1.2) except $\Delta^2 w$ and Δv into their right sides and considering these equations as elliptic ones on $\Omega^\bullet$ for fixed $t \in (\delta, T - \delta)$ from interior-type Schauder estimates for equations in variational form [1], we conclude that $(w_\delta, v_\delta) \in C(\delta, T - \delta;\ H_4(\Omega^\bullet) \times H_2(\Omega^\bullet))$. In addition, all terms of these equations involving ∂_t are in $L^2(Q_{\epsilon/2})$. By Theorem 1.3 we have $(w_\delta, v_\delta) = 0$ in Q_ϵ. Letting $\delta \to 0$, we obtain the same conclusion for w, v in any Q_ϵ, which concludes the proof. $\square$

This method of the proof implies conditional Hölder-type stability estimates for the Cauchy problem (1.1)–(1.3) (compare with [10, section 3.2]).

We think that it would be interesting to use the additional large parameter in Carleman estimates for boundary value problems considered by Tataru in [17] and to apply this method for a full (vector) system of thermoelasticity.

Acknowledgments. The author is grateful to Professor Irena Lasiecka who stimulated his work on uniqueness of the continuation for the von Karman and thermoelasticity systems, to Dr. Matthias Eller who helped to eliminate a number of flaws in the proofs, and to the referees for careful reading of the manuscript and helpful suggestions.

REFERENCES

[1] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for the solutions of elliptic partial differential equations satisfying general boundary conditions*, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

[2] G. Avalos and I. Lasiecka, *Boundary controllability of thermoelastic plates with free boundary conditions*, SIAM J. Control Optim., 38 (2000), pp. 337–383.

[3] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[4] M. Eller, V. Isakov, G. Nakamura, and D. Tataru, *Uniqueness and stability in the Cauchy problem for Maxwell's and elasticity systems*, College de France Seminar, Vol. 14, D. Cioranescu and J.-L. Lions, eds., Res. Notes in Math., Chapman and Hall/CRC, to appear.

[5] L. Hörmander, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, Heidelberg, New York, 1963.

[6] L. Hörmander, *The Analysis of Linear Partial Differential Operators*, Springer-Verlag, Berlin, Heidelberg, New York, 1983–1985.

[7] V. Isakov, *A nonhyperbolic Cauchy problem for $\square_b \square_c$ and its applications to elasticity theory*, Comm. Pure Appl. Math., 39 (1986), pp. 744–769.

[8] V. Isakov, *Carleman estimates in an anisotropic case and applications*, J. Differential Equations, 105 (1993), pp. 217–239.

[9] V. Isakov, *On uniqueness in the Cauchy problem with multiple characteristics*, J. Differential Equations, 133 (1997), pp. 134–147.

[10] V. Isakov, *Inverse problems for partial differential equations*, Appl. Math. Sci. 127, Springer-Verlag, New York, 1997.

[11] F. John, *Continuous dependence on data for solutions of partial differential equations with a prescribed bound*, Comm. Pure Appl. Math., 13 (1960), pp. 551–585.

[12] J. Lagnese, *The reachability problem for thermoelastic plates*, Arch. Ration. Mech. Anal., 112 (1990), pp. 223–267.

[13] G. Lebeau and E. Zuazua, *Null-controllability of system of linear thermoelasticity*, Arch. Ration. Mech. Anal., 141 (1998), pp. 297–329.

[14] L. Nirenberg, *Uniqueness of Cauchy problems for differential equations with constant leading coefficients*, Comm. Pure Appl. Math., 10 (1957), pp. 89–105.

[15] L. Robbiano and C. Zuily, *Uniqueness in the Cauchy problem for operators with partially holomorphic coefficients*, Invent. Math., 131 (1998), pp. 493–539.

[16] D. Tataru, *Unique continuation for solutions of PDEs: Between Hörmander's theorem and Holmgren's theorem*, Comm. Partial Differential Equations, 20 (1995), pp. 855–894.

[17] D. Tataru, *Carleman estimates and unique continuation for solutions to boundary value problems*, J. Math. Pures Appl., 75 (1996), pp. 367–408.

SIAM J. MATH. ANAL.
Vol. 33, No. 3, pp. 523–544

ASYMPTOTIC PROFILES OF NONSTATIONARY INCOMPRESSIBLE NAVIER–STOKES FLOWS IN THE WHOLE SPACE[*]

YOSHIKO FUJIGAKI[†] AND TETSURO MIYAKAWA[‡]

Dedicated to Professor Kiyoshi Asano

Abstract. Asymptotic profiles are deduced for weak and strong solutions of the incompressible Navier–Stokes equations in the whole space. It is shown that if the initial velocity satisfies a specific moment condition, the corresponding solution behaves like the first-order spatial derivatives of the heat kernel. Higher-order asymptotics are also deduced in case the initial data admit vector potentials with spatial decay of order $-n$. We further note that the results are not optimal and suggest by means of an example that there exist solutions with a more rapid (space-time) decay property if we require certain symmetry conditions to the initial data.

Key words. incompressible Navier–Stokes equations, initial value problem, asymptotic profiles, moment estimates

AMS subject classifications. 35Q30, 76D05

PII. S0036141000367072

1. Introduction.

We consider the Navier–Stokes system in $\mathbb{R}^n$, $n \geq 2$, which will be treated in this paper in the form of the integral equation

$$u(t) = e^{-tA}a - \int_0^t \nabla \cdot e^{-(t-s)A} P(u \otimes u)(s)ds, \tag{1.1}$$

and we discuss asymptotic properties of weak and strong solutions. Here, $u = (u_1, \ldots, u_n)$ is unknown velocity, $a = (a_1, \ldots, a_n)$ is a given initial velocity, $\nabla = (\partial_1, \ldots, \partial_n)$ with $\partial_j = \partial/\partial x_j$, $A = -\Delta$ is the Laplacian, $\{e^{-tA}\}_{t\geq 0}$ is the heat semigroup, $P = (P_{jk})$ is the Fujita–Kato bounded projection [5] onto the spaces of solenoidal vector fields, and

$$(\nabla \cdot e^{-tA} P(u \otimes u))_j = \sum_{k,\ell=1}^{n} \partial_\ell E_t * P_{jk} * (u_\ell u_k),$$

where $E_t = (4\pi t)^{-\frac{n}{2}} \exp(-\frac{|x|^2}{4t})$ is the heat kernel and $*$ is the convolution of distributions over $\mathbb{R}^n$. In this paper we always assume that the initial data a are solenoidal, i.e., $\nabla \cdot a = 0$, and satisfy

$$\int (1 + |y|)|a(y)|dy < \infty. \tag{1.2}$$

[*] Received by the editors January 31, 2000; accepted for publication (in revised form) February 14, 2001; published electronically September 5, 2001.
http://www.siam.org/journals/sima/33-3/36707.html

[†] Graduate School of Science and Technology, Kobe University, Rokko, Kobe 657-8501, Japan (fujigaki@math.sci.kobe-u.ac.jp). This author's work was partially supported by the Japan Scholarship Foundation.

[‡] Department of Mathematics, Faculty of Science, Kobe University, Rokko, Kobe 657-8501, Japan (miyakawa@math.sci.kobe-u.ac.jp). This author's work was partially supported by Grant-in-Aid for Scientific Research, 10440045 and 12874027, from the Japan Society for Promotion of Science.

Here, and in what follows, integration will be performed on $\mathbb{R}^n$ unless otherwise specified. The assumption (1.2) implies $a \in \boldsymbol{L}^1$; so the condition $\nabla \cdot a = 0$ ensures (see [7, 8])

$$\int a(y)dy = 0. \tag{1.3}$$

Most literature on the nonstationary Navier–Stokes system deals with solutions as curves in some L^p-like function spaces and only a few results are available concerning the space-time asymptotic behavior of the solutions. The weak solutions are treated as curves in the space $\boldsymbol{L}^2$ and the existence for all $t > 0$ is established for all initial data in $\boldsymbol{L}^2$, although its uniqueness still remains open when $n \geq 3$. As for the strong solutions, it is known (see [5, 8]) that a unique strong solution $u(t)$ exists for all $t > 0$ in general $\boldsymbol{L}^q$-spaces, satisfying

$$\|u(t)\|_q \leq Ct^{-\frac{1}{2}-\frac{n}{2}(1-\frac{1}{q})}, \quad \|\nabla u(t)\|_q \leq Ct^{-1-\frac{n}{2}(1-\frac{1}{q})} \qquad (1 \leq q \leq \infty, \quad t > 0), \tag{1.4}$$

if a is in $\boldsymbol{L}^n \cap \boldsymbol{L}^1$, is small in $\boldsymbol{L}^n$, and satisfies (1.2). Hereafter, $\| \cdot \|_r$ denotes the L^r-norm. On the other hand, the second author proved in [9] that given γ with $1 \leq \gamma \leq n+1$, problem (1.1) admits, under suitable assumptions on a, a unique strong solution u such that

$$|u(x,t)| \leq C(1+|x|)^{-\alpha}(1+t)^{-(\gamma-\alpha)/2} \qquad \text{for all } \alpha \text{ with } 0 \leq \alpha \leq \gamma, \tag{1.5}$$

and improved a similar result of Takahashi [17]. Inspired by [9], we deduce in this paper another kind of space-time asymptotic profile of u for weak and strong solutions. To be more precise, we first show that if a satisfies (1.2), the weak and strong solutions given by the standard method admit, as $t \to \infty$, an asymptotic expansion of the first order in terms of the spatial derivatives of Gaussian-like functions. As for the strong solutions, our result improves that of Carpio [1] which deduces the first-order asymptotics of two kinds, one in $\mathbb{R}^3$ and the other in $\mathbb{R}^2$. Our proof shows that one and the same result holds in all space dimensions $n \geq 2$. Moreover, contrary to [1], our argument requires neither the theory of Hardy spaces nor the Calderón–Zygmund kernels but utilizes only Taylor's formula for smooth functions and elementary results on the Fourier transform.

We next consider the strong solutions satisfying (1.5) in $\mathbb{R}^n$, $n \geq 2$, and the weak solutions in $\mathbb{R}^3$ and $\mathbb{R}^4$. We show that these solutions admit a higher-order asymptotic expansion in terms of the space-time derivatives of Gaussian-like functions if the initial data satisfy appropriate moment conditions. We prove this result with the aid of (an improvement of) the estimates for L^2-moments of solutions as given in [3, 14].

In section 2 we state our main results after some preliminaries. In particular, we prove there that the kernel function of the operator $\nabla \cdot e^{-tA}P$ belongs to $L^1 \cap L^\infty$ together with its derivatives and behaves like the first-order derivative of the heat kernel. Due to this nice property of the kernel function, we can avoid the use of Hardy spaces and Calderón–Zygmund kernels.

In section 3 we deduce the asymptotics for the linear term $e^{-tA}a$ and in section 4 we prove the first-order asymptotics for weak and strong solutions whose initial data satisfy (1.2). The weak solutions treated in this paper are those satisfying the energy inequality

$$\|u(t)\|_2^2 + 2\int_0^t \|\nabla u\|_2^2 ds \leq \|a\|_2^2 \qquad \text{for all } t \geq 0 \tag{1.6}$$

and the decay estimate

$$\|u(t)\|_2 \leq C(1+t)^{-\frac{n+2}{4}}. \tag{1.7}$$

It is now well known (see [4, appendix] and [18]) that such weak solutions exist in all space dimensions $n \geq 2$ whenever $a \in \boldsymbol{L}^2$, $\nabla \cdot a = 0$, and a satisfies (1.2). We shall deduce the first-order asymptotics for such weak solutions by slightly modifying the argument given for strong solutions.

In section 5 we deduce higher-order asymptotics for strong solutions satisfying (1.5) with $\gamma = n+1$ and for weak solutions whose initial data satisfy (1.2) and some additional moment conditions. It should be noticed here that our strong solutions admit the asymptotic expansion up to (and including) order n, the space dimension, while our weak solutions admit the same expansion only up to (and including) order $n-1$. This is because of the difference between the L^2-moment estimates which are satisfied by our strong or weak solutions (see (2.7) and (2.8) below). We also note that the moment estimate (2.8) for weak solutions is proved only when $n = 3, 4$, so our higher-order asymptotic result for weak solutions deals only with this case.

In section 6 we examine the solutions in $\mathbb{R}^2$ which decay exponentially in time and are rapidly decreasing in the spatial direction, and we give a slightly refined version of a result of Schonbek [12] (see also [15]). The result shows that our asymptotic results are by no means optimal. We refer the reader to [10, 11, 12, 13, 15] for the problem of the optimality of decay rates of solutions in connection with the classes of corresponding initial data.

In deducing the higher-order expansion for strong solutions, we have treated in this paper only those solutions which satisfy (1.5) with $\gamma = n+1$. It would be an interesting problem to find space-time asymptotic profiles in the case $1 \leq \gamma < n+1$. We also note that nothing is known about the space-time behavior for weak solutions if we drop assumption (1.2) on initial data.

Finally we note that our first-order asymptotic result is extended to weak and strong solutions of the Navier–Stokes system in the half-space $\mathbb{R}^n_+$. The result clarifies a difference of the behavior between the Navier–Stokes flows in $\mathbb{R}^n$ and $\mathbb{R}^n_+$ which is caused by the presence of the boundary of $\mathbb{R}^n_+$. The details are given in [2].

2. Preliminaries and the results.

We first recall that the Fujita–Kato projection P onto the solenoidal fields has the kernel function $P(x) = (P_{jk}(x))_{j,k=1}^n$ with the Fourier transform

$$\widehat{P}_{jk}(\xi) \equiv \int e^{-ix\cdot\xi} P_{jk}(x)dx = \delta_{jk} + \frac{i\xi_j i\xi_k}{|\xi|^2} \qquad \left(i = \sqrt{-1}, \quad x \cdot \xi = \sum_{j=1}^n x_j \xi_j \right).$$

Therefore, $\partial_\ell e^{-tA}P = F_\ell = (F_{\ell,jk})_{j,k=1}^n$ with

$$\widehat{F}_{\ell,jk}(\xi,t) = i\xi_\ell e^{-t|\xi|^2}\left(\delta_{jk} + \frac{i\xi_j i\xi_k}{|\xi|^2}\right) \equiv \widehat{F}^1_{\ell,jk}(\xi,t) + \widehat{F}^2_{\ell,jk}(\xi,t).$$

Thus, denoting the heat kernel by

$$E_t(x) = (4\pi t)^{-\frac{n}{2}} \exp(-\tfrac{|x|^2}{4t}),$$

we easily see that $F^1_{\ell,jk}(x,t) = (\partial_\ell E_t)(x)\delta_{jk}$, and so, writing $\partial_x^\beta = \partial_1^{\beta_1} \cdots \partial_n^{\beta_n}$ for any multi-index $\beta = (\beta_1, \ldots, \beta_n)$ of nonnegative integers,

$$\|\partial_t^p \partial_x^\beta F^1_{\ell,jk}(\cdot,t)\|_q \leq C_q t^{-\frac{1+|\beta|+2p}{2} - \frac{n}{2}(1-\frac{1}{q})} \qquad (1 \leq q \leq \infty).$$

To evaluate $F^2_{\ell,jk}$, we invoke the relation $|\xi|^{-2} = \int_0^\infty e^{-s|\xi|^2} ds$ and get

$$\widehat{F}^2_{\ell,jk}(\xi,t) = i\xi_\ell i\xi_j i\xi_k \int_t^\infty e^{-s|\xi|^2} ds \qquad \text{so that} \quad F^2_{\ell,jk}(x,t) = \int_t^\infty \partial_\ell\partial_j\partial_k E_s(x) ds.$$

From this we easily obtain

$$\|\partial_t^p \partial_x^\beta F^2_{\ell,jk}(\cdot,t)\|_q \le C_q t^{-\frac{1+|\beta|+2p}{2}-\frac{n}{2}(1-\frac{1}{q})} \qquad (1 \le q \le \infty).$$

Combining this with the estimate for $F^1_{\ell,jk}$ gives

$$\|\partial_t^p \partial_x^\beta F_{\ell,jk}(\cdot,t)\|_q \le C_q t^{-\frac{1+|\beta|+2p}{2}-\frac{n}{2}(1-\frac{1}{q})} \qquad (1 \le q \le \infty). \tag{2.1}$$

In this paper we employ the summation convention for repeated indices. Our results are then stated as follows.

THEOREM 2.1.

(i) *Let $a \in \boldsymbol{L}^n \cap \boldsymbol{L}^1$ be solenoidal and satisfy* (1.2). *Let $u = (u_1, \dots, u_n)$ be the corresponding strong solution of* (1.1) *which exists for all $t \ge 0$ if a is small in $\boldsymbol{L}^n$. Then for $1 \le q \le \infty$ and $j = 1, \dots, n$, we have*

$$\begin{aligned} \lim_{t\to\infty} t^{\frac{1}{2}+\frac{n}{2}(1-\frac{1}{q})} \Big\| u_j(t) + (\partial_k E_t)(\cdot) \int y_k a_j(y) dy \\ + F_{\ell,jk}(\cdot,t) \int_0^\infty \int (u_\ell u_k)(y,s) dy ds \Big\|_q = 0. \end{aligned} \tag{2.2}$$

(ii) *For every $a \in \boldsymbol{L}^2$ which is solenoidal and satisfies* (1.2), *there exists a weak solution u which admits the expansion* (2.2) *with $1 \le q \le 2$. The result below concerns higher-order asymptotics of weak and strong solutions.*

THEOREM 2.2.

(iii) *Let a satisfy the assumption of Theorem* 2.1(i) *and the following additional conditions:*

$$\begin{aligned} &\int |y|^m |a(y)| dy < \infty, && |a(y)| \le c_0(1+|y|)^{-n-1}, && \\ &a_j = \sum_{k=1}^n \partial_k b_{jk}, && |b_{jk}(y)| \le c_0(1+|y|)^{-n}, && b_{jk} \in L^1, \end{aligned} \tag{2.3}$$

for some integer m such that $1 \le m \le n$. If $c_0 > 0$ and the norms $\|b_{jk}\|_1$ are small, there exists a global strong solution u which satisfies (1.5) *with $\gamma = n+1$. Furthermore, for $1 \le q \le \infty$ and $j = 1, \dots, n$, we have*

$$\begin{aligned} \lim_{t\to\infty} t^{\frac{m}{2}+\frac{n}{2}(1-\frac{1}{q})} \Big\| u_j(t) - \sum_{1\le|\alpha|\le m} \frac{(-1)^{|\alpha|}}{\alpha!} (\partial_x^\alpha E_t)(\cdot) \int y^\alpha a_j(y) dy \\ + \sum_{|\beta|+2p\le m-1} \frac{(-1)^{|\beta|+p}}{p!\beta!} (\partial_t^p \partial_x^\beta F_{\ell,jk})(\cdot,t) \\ \times \int_0^\infty \int s^p y^\beta (u_\ell u_k)(y,s) dy ds \Big\|_q = 0. \end{aligned} \tag{2.4}$$

(iv) *Let $n = 3, 4$, and suppose that*

$$\int (1+|y|)^{n-1}|a(y)|dy < \infty, \qquad \int (1+|y|)^n|a(y)|^2 dy < \infty. \tag{2.5}$$

Then there exists a weak solution u satisfying (2.4) *for* $1 \le q \le 2$ *and* $1 \le m \le n-1$.

Remarks.

(i) Notice that (1.4) implies

$$\|u(s)\|_2^2 \le C(1+s)^{-1-\frac{n}{2}}, \tag{2.6}$$

so the last integral in (2.2) is finite. On the other hand, (1.2) and (1.3) together imply

$$\|e^{-tA}a\|_2^2 \le C(1+t)^{-1-\frac{n}{2}}.$$

So a result of Wiegner [18] ensures the existence of a weak solution u satisfying (2.6).

(ii) Theorem 2.1 improves an asymptotic result of Carpio [1] in the following sense. First, the result of [1] ignores the vanishing of the average (1.3) and so contains the trivial term $E_t(x)\int a(y)dy \equiv 0$. Second, [1] deals only with the case discussed in assertion (i) of Theorem 2.1, and the results given there are incomplete in the two-dimensional case.

(iii) The existence of a strong solution treated in Theorem 2.2(iii) is proved in [9], and convergence of the integrals in the second sum of (2.4) is ensured by the estimate

$$\int |y|^m|u(y,s)|^2 dy \le C(1+s)^{-\frac{n-m}{2}-1} \qquad (0 \le m \le n+1). \tag{2.7}$$

This estimate will be proved in section 5.

(iv) The proofs of Theorems 2.1 and 2.2 will be carried out in almost the same way for weak and strong solutions. They differ only in estimating the nonlinear convolution integral of (1.1) in a neighborhood of $s = t$. The restriction $m \le n-1$ in Theorem 2.2(iv) arises from the fact that for weak solutions we know only the estimate

$$\int |y|^m|u(y,s)|^2 dy \le C(1+s)^{-(1+\frac{n}{2})(1-\frac{m}{n})} \qquad (0 \le m \le n, \quad n = 3, 4), \tag{2.8}$$

which is weaker than (2.7). This estimate is due to [3, 14], and a detailed proof will be given in the appendix for the reader's convenience. Since (2.8) seems to be valid for general weak solutions only when $n = 3, 4$, Theorem 2.2(iv) would be valid only for $n = 3$, 4. This point will be discussed at the end of section 5.

It should be emphasized here that Theorems 2.1 and 2.2 are by no means optimal. Indeed, the following result is known.

THEOREM 2.3. *If $n = 2$, a solution u exists satisfying*

$$\|u(t)\|_q \le C_q e^{-\gamma_q t} \qquad \textit{and} \qquad |u(x,t)| \le C_m e^{-\gamma t}(1+|x|)^{-m}$$

for all $1 \le q \le \infty$ and $m = 0, 1, 2, \dots$, with some positive constants C_q, C_m, γ, γ_q, and γ_m.

This result is proved in [12] for $2 \le q \le \infty$ and $n = 2$, and it is extended in [15] to the case when n is even. Our Theorem 2.3 covers the case $1 \le q < 2$ and contains a pointwise decay result. We give a detailed proof in section 6 for the reader's convenience.

In [11, 12, 13] Schonbek and in [15] Schonbek, Schonbek, and Süli discuss the problem of finding lower bounds of rates of decay in time for weak solutions to the Navier–Stokes system which do not belong to the class of solutions as described in Theorem 2.3. We note that Theorem 2.1(ii) can be applied to characterizing weak solutions satisfying the lower bound estimate $\|u(t)\|_2 \geq ct^{-\frac{n+2}{4}}$ for large $t > 0$. The details are given in [10].

3. Asymptotics for the linear term. This section proves the following theorem.

THEOREM 3.1. *Suppose a is solenoidal and satisfies*

$$\int (1+|y|)^m |a(y)| dy < \infty \tag{3.1}$$

for an integer $m \geq 1$. Then for $1 \leq q \leq \infty$,

$$\lim_{t\to\infty} t^{\frac{m}{2}+\frac{n}{2}(1-\frac{1}{q})} \left\| e^{-tA}a - \sum_{1\leq|\alpha|\leq m} \frac{(-1)^{|\alpha|}}{\alpha!} (\partial_x^\alpha E_t)(\cdot) \int y^\alpha a(y)dy \right\|_q = 0. \tag{3.2}$$

Proof. Recall that (see [7, 8]) since a is solenoidal and integrable, it satisfies (1.3). Thus, applying Taylor's formula gives

$$\begin{aligned}(e^{-tA}a)(x) &\equiv \int E_t(x-y)a(y)dy = \int [E_t(x-y) - E_t(x)]a(y)dy \\ &= \sum_{1\leq|\alpha|\leq m-1} \frac{(-1)^{|\alpha|}}{\alpha!} (\partial_x^\alpha E_t)(x) \int y^\alpha a(y)dy + \int R_m(x,y)a(y)dy,\end{aligned}$$

where

$$\begin{aligned}R_m(x,y) &= \frac{1}{(m-1)!}\int_0^1 (1-\theta)^{m-1} \left(\frac{d}{d\theta}\right)^m E_t(x-y\theta)d\theta \\ &= \sum_{|\alpha|=m} \frac{(-1)^{|\alpha|}}{\alpha!}(\partial_x^\alpha E_t)(x)y^\alpha \\ &\quad + \frac{(-1)^m}{(m-1)!}\int_0^1 (1-\theta)^{m-1} \sum_{|\alpha|=m} \frac{m!}{\alpha!}[(\partial_x^\alpha E_t)(x-y\theta) - (\partial_x^\alpha E_t)(x)]y^\alpha d\theta.\end{aligned}$$

Therefore, via the change of variables $xt^{-\frac{1}{2}} \to x$ we obtain

$$\begin{aligned}&\left\| e^{-tA}a - \sum_{1\leq|\alpha|\leq m} \frac{(-1)^{|\alpha|}}{\alpha!} (\partial_x^\alpha E_t)(\cdot) \int y^\alpha a(y)dy \right\|_q \\ &\qquad \leq C_m t^{-\frac{m}{2}-\frac{n}{2}(1-\frac{1}{q})} \sum_{|\alpha|=m} \int_0^1 \int \varphi_t(y,\theta)|y|^m |a(y)|dyd\theta,\end{aligned}$$

where

$$\varphi_t(y,\theta) = \sum_{|\alpha|=m} \|(\partial_x^\alpha E_1)(\cdot - y\theta t^{-\frac{1}{2}}) - (\partial_x^\alpha E_1)(\cdot)\|_q.$$

This function is bounded in t, θ, and y, and we have

$$\lim_{t\to\infty} \varphi_t(y,\theta) = 0 \qquad \text{for fixed } y \text{ and } \theta. \tag{3.3}$$

Since $|y|^m|a(y)|$ is integrable on $\mathbb{R}^n$ by (3.1), the dominated convergence theorem yields

$$\lim_{t\to\infty} \int_0^1 \int \varphi_t(y,\theta)|y|^m|a(y)|dyd\theta = 0.$$

This implies (3.2) and so the proof of Theorem 3.1 is complete. □

Remark. Convergence (3.3) is valid for $q = \infty$ since the function $\partial_x^\alpha E_1$ is bounded and uniformly continuous on $\mathbb{R}^n$.

4. Proof of Theorem 2.1. Let

$$\begin{aligned} w(t) &= (w_1(t),\dots,w_n(t)) = -\int_0^t \nabla\cdot e^{-(t-s)A}P(u\otimes u)(s)ds \\ &= -\left(\int_0^t F_{\ell,jk}(t-s) * (u_\ell u_k)(s)ds\right)_{j=1}^n . \end{aligned} \tag{4.1}$$

Due to Theorem 3.1, it suffices to prove the following theorem.

THEOREM 4.1.

(i) *Under the assumption of Theorem* 2.1(i), *we have*

$$\lim_{t\to\infty} t^{\frac12+\frac n2(1-\frac1q)} \left\| w_j(t) + F_{\ell,jk}(\cdot,t)\int_0^\infty\int (u_\ell u_k)(y,s)dyds\right\|_q = 0 \tag{4.2}$$

for all $1 \le q \le \infty$ *and* $j = 1,\dots,n$.

(ii) *Under the assumption of Theorem* 2.1(ii), *the weak solution* u *satisfies* (4.2) *for all* $1 \le q \le 2$ *and* $j = 1,\dots,n$.

Proof. We write (4.1) as

$$w_j(t) = -\left(\int_0^{t/2} + \int_{t/2}^t\right) F_{\ell,jk}(t-s) * (u_\ell u_k)(s)dyds \equiv J_1 + J_2.$$

Direct calculation gives

$$\begin{aligned} w_j(t) &+ F_{\ell,jk}(x,t)\int_0^\infty\int (u_\ell u_k)dyds \\ &= F_{\ell,jk}(x,t)\int_{t/2}^\infty\int (u_\ell u_k)dyds \\ &\quad - \int_0^{t/2}\int [F_{\ell,jk}(x-y,t-s) - F_{\ell,jk}(x,t-s)](u_\ell u_k)dyds \\ &\quad + \int_0^{t/2}\int\int_0^1 s(\partial_t F_{\ell,jk})(x,t-s\tau)(u_\ell u_k)dydsd\tau + J_2 \\ &\equiv J_{11} + J_{12} + J_{13} + J_2. \end{aligned}$$

We see from (2.1) and (2.6) that

$$t^{\frac12+\frac n2(1-\frac1q)}\|J_{11}\|_q \le C_q \int_{t/2}^\infty \|u(s)\|_2^2 ds \le Ct^{-\frac n2} \to 0 \qquad \text{as } t\to\infty \tag{4.3}$$

for all $1 \le q \le \infty$. Similarly, applying (2.1) and (2.6) gives

$$\|J_{13}\|_q \le C_q \int_0^1 \int_0^{t/2} s(t-s\tau)^{-\frac{3}{2}-\frac{n}{2}(1-\frac{1}{q})} \|u(s)\|_2^2 ds d\tau$$

$$\le C_q t^{-\frac{3}{2}-\frac{n}{2}(1-\frac{1}{q})} \int_0^{t/2} s\|u(s)\|_2^2 ds \le C_q t^{-\frac{3}{2}-\frac{n}{2}(1-\frac{1}{q})} \int_0^t (1+s)^{-\frac{n}{2}} ds$$

so that, for all $1 \le q \le \infty$,

$$t^{\frac{1}{2}+\frac{n}{2}(1-\frac{1}{q})} \|J_{13}\|_q \le Ct^{-1} \int_0^t (1+s)^{-\frac{n}{2}} ds \to 0 \qquad \text{as } t \to \infty. \tag{4.4}$$

Next, we write $F_{\ell,jk}(x,t) = t^{-\frac{1+n}{2}} K(xt^{-\frac{1}{2}})$ in terms of a smooth, bounded, integrable, and uniformly continuous function K. Applying Minkowski's inequality for the integral yields, after a change of variables,

$$\|J_{12}\|_q \le C_q \int_0^{t/2} \int (t-s)^{-\frac{1}{2}-\frac{n}{2}(1-\frac{1}{q})} \|K(\cdot - y(t-s)^{-\frac{1}{2}}) - K(\cdot)\|_q |u(y,s)|^2 dy ds$$

$$\le C_q t^{-\frac{1}{2}-\frac{n}{2}(1-\frac{1}{q})} \int_0^{t/2} \int \|K(\cdot - y(t-s)^{-\frac{1}{2}}) - K(\cdot)\|_q |u(y,s)|^2 dy ds.$$

Therefore,

$$t^{\frac{1}{2}+\frac{n}{2}(1-\frac{1}{q})} \|J_{12}\|_q \le C_q \int_0^{t/2} \int \varphi_t(y,s) |u(y,s)|^2 dy ds = C_q \int_0^{t/2} \psi_t(s) ds,$$

where $\varphi_t(y,s) = \|K(\cdot - y(t-s)^{-\frac{1}{2}}) - K(\cdot)\|_q$ and $\psi_t(s) = \int \varphi_t(y,s)|u(y,s)|^2 dy$. Note that $\varphi_t(y,s) \le C_q$, that $\varphi_t(y,s) \to 0$ as $t \to \infty$ for fixed y and s, and that $|u(y,s)|^2 dy$ is a finite measure on $\mathbb{R}^n$ for fixed s. The bounded convergence theorem now implies $\psi_t(s) \to 0$ as $t \to \infty$ for each fixed s. However, $\psi_t(s) \le C_q \|u(s)\|_2^2$, and the right-hand side is bounded and integrable over $[0,\infty)$ due to (2.6). Applying again the bounded convergence theorem gives

$$\lim_{t\to\infty} \int_0^M \psi_t(s) ds = 0 \qquad \text{for any fixed } M > 0. \tag{4.5}$$

Now, given $\varepsilon > 0$, choose $M > 0$ so that $\int_M^\infty \|u(s)\|_2^2 ds < \varepsilon$. Then for $t > 2M$, we have

$$\int_0^{t/2} \psi_t(s) ds \le \int_0^M \psi_t(s) ds + C_q \int_M^\infty \|u(s)\|_2^2 ds \le \int_0^M \psi_t(s) ds + C_q \varepsilon.$$

This, together with (4.5), gives $\lim_{t\to\infty} \int_0^{t/2} \psi_t(s) ds = 0$; and we have deduced

$$\lim_{t\to\infty} t^{\frac{1}{2}+\frac{n}{2}(1-\frac{1}{q})} \|J_{12}\|_q = 0 \qquad \text{for all } 1 \le q \le \infty. \tag{4.6}$$

Observe that we have so far invoked only (2.6) for estimating u, so (4.3), (4.4), and (4.6) hold for both of the weak and strong solutions.

We next estimate J_2. It is here where we have to deal with weak and strong solutions separately. Consider first the strong solutions. By (1.4) we get

$$\|J_2\|_q \le \int_{t/2}^{t} \|F_{\ell,jk}(\cdot, t-s)\|_1 \|u(s)\|_{2q}^2 ds$$
$$\le C_q \int_{t/2}^{t} (t-s)^{-\frac{1}{2}}(1+s)^{-1-(n-\frac{n}{2q})} ds \le C_q (1+t)^{-\frac{1}{2}-(n-\frac{n}{2q})}$$

for all $1 \le q \le \infty$, and so

$$t^{\frac{n}{2}+\frac{n}{2}(1-\frac{1}{q})}\|J_2\|_q \le C_q(1+t)^{-\frac{1}{2}} \to 0 \qquad \text{as } t \to \infty. \tag{4.7}$$

This completes the proof of Theorem 4.1(i). □

Next consider the weak solutions. We first show that

$$\lim_{t\to\infty} t^{\frac{n}{2}}\|J_2\|_1 = 0. \tag{4.8}$$

We apply (2.1) and (2.6) to get

$$\|J_2\|_1 \le \int_{t/2}^{t} \|F(t-s)\|_1 \|u(s)\|_2^2 ds \le C \int_{t/2}^{t} (t-s)^{-\frac{1}{2}} s^{-1-\frac{n}{2}} ds \le Ct^{-\frac{n+1}{2}},$$

so that $t^{\frac{n}{2}}\|J_2\|_1 \le Ct^{-\frac{1}{2}} \to 0$ as $t \to \infty$, which proves (4.8). Second, we show that

$$\lim_{t\to\infty} t^{\frac{n}{2}+\frac{n}{4}}\|J_2\|_2 = 0. \tag{4.9}$$

The argument below is due to [4] (see also [11, 18]). Let

$$v(t) = -\int_{\tau}^{t} F(t-s) * (u \otimes u)(s) ds = u(t) - e^{-(t-\tau)A}u(\tau)$$

with $0 < \tau < t$, and assume that v is smooth. (This situation is realized if we replace u by approximate solutions u_N as given in [4, 11].) Then v solves the initial value problem

$$\partial_t v + Av = -P(u \cdot \nabla u) \quad (t > \tau), \qquad v(\tau) = 0.$$

Multiplying the above equation by $2v$ and integrating by parts gives, since $(u\cdot\nabla v, v) = 0$,

$$\partial_t \|v\|_2^2 + 2\|A^{1/2}v\|_2^2 = -2(u\cdot\nabla u, v) = 2(u\cdot\nabla v, u) = 2(u\cdot\nabla v, u_0),$$

where $u_0(t) = e^{-(t-\tau)A}u(\tau)$. By the standard L^p-L^q estimates for e^{-tA} and (2.6), we get

$$\|u_0(t)\|_\infty \le C(t-\tau)^{-\frac{n}{4}}\|u(\tau)\|_2 \le C(t-\tau)^{-\frac{n}{4}}\tau^{-\frac{n}{4}-\frac{1}{2}}.$$

Since $\|\nabla v\|_2 = \|A^{1/2}v\|_2$ and $\|u\|_2 \le \|a\|_2$, we have

$$2|(u\cdot\nabla v, u_0)| \le C\|u\|_2\|\nabla v\|_2\|u_0\|_\infty = C\|A^{1/2}v\|_2\|u\|_2\|u_0\|_\infty$$
$$\le C\|A^{1/2}v\|_2(t-\tau)^{-\frac{n+1}{2}}\tau^{-\frac{n}{4}-\frac{1}{2}}$$
$$\le \|A^{1/2}v\|_2^2 + C(t-\tau)^{-n-1}\tau^{-\frac{n}{2}-1},$$

which implies

$$\partial_t\|v\|_2^2 + \|A^{1/2}v\|_2^2 \le C(t-\tau)^{-n-1}\tau^{-\frac{n}{2}-1}.$$

Let $\{E_\lambda\}_{\lambda\ge 0}$ be the spectral measure associated with the positive self-adjoint operator A. Applying $\|A^{1/2}v\|_2^2 \ge \varrho(\|v\|_2^2 - \|E_\varrho v\|_2^2)$ yields

$$\partial_t\|v\|_2^2 + \varrho\|v\|_2^2 \le \varrho\|E_\varrho v\|_2^2 + C(t-\tau)^{-n-1}\tau^{-\frac{n}{2}-1}.$$

However, we know that (see [4, 11, 18])

$$\|E_\varrho v\|_2^2 \le C\varrho^{\frac{n+2}{2}}\left(\int_\tau^t \|u\|_2^2 ds\right)^2$$

and so

$$\partial_t\|v\|_2^2 + \varrho\|v\|_2^2 \le C\varrho^{\frac{n+4}{2}}\left(\int_\tau^t \|u\|_2^2 ds\right)^2 + C(t-\tau)^{-n-1}\tau^{-\frac{n}{2}-1}.$$

Here we put $\varrho = m(t-\tau)^{-1}$, $m>0$; then we multiply both sides by $(t-\tau)^m$ to get

$$\partial_t((t-\tau)^m\|v\|_2^2) \le C(t-\tau)^{m-\frac{n}{2}-2}\left(\int_\tau^t \|u\|_2^2 ds\right)^2 + C(t-\tau)^{m-n-1}\tau^{-\frac{n}{2}-1}.$$

Choosing m so that $m > n/2+2$ and $m > n+1$, we obtain

$$\begin{aligned}\|v(t)\|_2^2 &\le C(t-\tau)^{-m}\int_\tau^t (s-\tau)^{m-\frac{n}{2}-2}\left(\int_\tau^s \|u\|_2^2 d\sigma\right)^2 ds + C(t-\tau)^{-n}\tau^{-\frac{n}{2}-1}\\ &\le C(t-\tau)^{-2-\frac{n}{2}}\int_\tau^t\left(\int_\tau^s \|u\|_2^2 d\sigma\right)^2 ds + C(t-\tau)^{-n}\tau^{-1-\frac{n}{2}}.\end{aligned}$$

Inserting $\tau = t/2$ yields $v(t) = J_2$, and so

$$t^{n+\frac{n}{2}}\|J_2\|_2^2 \le Ct^{n-1}\left(\int_{t/2}^\infty \|u\|_2^2 ds\right)^2 + Ct^{-1} \le Ct^{-1} \to 0 \qquad \text{as } t\to\infty.$$

This proves (4.9). Interpolating between (4.8) and (4.9) now gives

$$\lim_{t\to\infty} t^{\frac{n}{2}+\frac{n}{2}(1-\frac{1}{q})}\|J_2\|_q = 0 \qquad \text{for all } 1\le q\le 2. \tag{4.10}$$

This completes the proof of Theorem 4.1(ii). □

5. Proof of Theorem 2.2. This section proves Theorem 2.2. Recall that, as shown in [9], our strong solutions satisfy pointwise estimate (1.5) with $\gamma = n+1$, i.e.,

$$|u(x,t)| \le C_\alpha(1+|x|)^{\alpha-n-1}(1+t)^{-\alpha/2} \qquad \text{for all } 0\le\alpha\le n+1. \tag{5.1}$$

Choosing $\alpha = n+1$ and then $\alpha = 1$, we get

$$\|u(t)\|_\infty \le C(1+t)^{-\frac{1+n}{2}}, \qquad \|u(t)\|_{1,w} \le C(1+t)^{-\frac{1}{2}}, \tag{5.2}$$

where $\|\cdot\|_{1,w}$ is the quasi norm of the weak L^1-space (see [16]). Hence, we get (2.6), i.e.,

$$\|u(t)\|_2 \le C\|u(t)\|_\infty^{1/2}\|u(t)\|_{1,w}^{1/2} \le C(1+t)^{-\frac{n+2}{4}}.$$

Using (1.2), (1.3), (2.1), and (2.6), we can estimate (1.1) to get $\|u(t)\|_1 \le C(1+t)^{-\frac{1}{2}}$. Combining this with the first estimate of (5.2), we conclude that

$$\|u(t)\|_q \le C(1+t)^{-\frac{1}{2}-\frac{n}{2}(1-\frac{1}{q})} \qquad \text{for all } 1 \le q \le \infty. \tag{5.3}$$

We also invoke estimate (2.7) for our strong solutions, i.e.,

$$\int |y|^m |u(y,s)|^2 dy \le C_m(1+s)^{-1-\frac{n-m}{2}} \qquad \text{for all } 0 \le m \le n+1. \tag{5.4}$$

This is deduced as follows. Note that (5.1) implies $|y|^{n+1}|u(y,s)| \le C$, and so

$$\int |y|^{n+1}|u(y,s)|^2 dy \le C\int |u(y,s)|dy \le C(1+s)^{-\frac{1}{2}}.$$

Combining this with (5.3) gives, via Hölder's inequality,

$$\int |y|^m|u(y,s)|^2 dy \le \left(\int |y|^{n+1}|u|^2 dy\right)^{\frac{m}{n+1}} \left(\int |u|^2 dy\right)^{1-\frac{m}{n+1}} \le C_m(1+s)^{-1-\frac{n-m}{2}}.$$

On the other hand, under the assumptions of Theorem 2.2(iv), we know (see [3, 14]) the existence of a weak solution u satisfying the energy inequality and (2.8), i.e.,

$$\int |y|^m |u(y,s)|^2 dy \le C(1+s)^{-(1+\frac{n}{2})(1-\frac{m}{n})} \qquad \text{for all } 0 \le m \le n. \tag{5.5}$$

The proof of (5.5) will be given in the appendix.

Now define the function $w(t)$ by (4.1). Since we have Theorem 3.1, in order to prove Theorem 2.2 we need only show the following.

THEOREM 5.1.

(i) *Under the asumption of Theorem* 2.2(iii), *we have*

$$\begin{aligned}\left\| w_j(t) + \sum_{|\beta|+2p\le m-1} \frac{(-1)^{|\beta|+p}}{p!\beta!}(\partial_x^\beta \partial_t^p F_{\ell,jk})(\cdot,t)\int_0^\infty\int s^p y^\beta (u_\ell u_k)dyds\right\|_q \\ = o(t^{\frac{m}{2}+\frac{n}{2}(1-\frac{1}{q})}) \qquad \textit{as } t\to\infty\end{aligned} \tag{5.6}$$

for all $1 \le q \le \infty$ *and all integers* m *such that* $1 \le m \le n$.

(ii) *Under the assumption of Theorem* 2.2(iv), *the function* w *satisfies* (5.6) *for all* $1 \le q \le 2$ *and all integers* m *such that* $1 \le m \le n-1$.

To prove Theorem 5.1, we again invoke the notation

$$w_j(t) = -\left(\int_0^{t/2} + \int_{t/2}^t\right) F_{\ell,jk}(t-s) * (u_\ell u_k)(s)ds \equiv J_1 + J_2.$$

The integral J_2 is already estimated in section 4, and we know that (4.7) holds for strong solutions and (4.10) for weak solutions. It thus suffices to find the desired expansion for J_1. As in section 4, J_1 is estimated in the same way for both weak and strong solutions.

We begin by noticing the following version of Taylor's formula.

LEMMA 5.2. *Let F denote any of the functions $F_{\ell,jk}$ and let $m \geq 1$ be an arbitrary integer. Then,*

$$F(x-y,t-s) = \sum_{|\beta|+2p\leq m-1} \frac{(-y)^\beta(-s)^p}{p!\beta!}(\partial_x^\beta \partial_t^p F)(x,t) + S^m.$$

Here, denoting $N_m = [(m-1)/2]$, the greatest integer in $(m-1)/2$,

$$\begin{aligned} S^m = &\sum_{|\beta|+2p=m-1,|\beta|\geq 2} |\beta| \int_0^1 (1-\theta)^{|\beta|-1} \frac{(-y)^\beta(-s)^p}{p!\beta!} \\ &\qquad\qquad \times[(\partial_x^\beta\partial_t^p F)(x-y\theta,t) - (\partial_x^\beta\partial_t^p F)(x,t)]d\theta \\ &+\frac{(-s)^{N_m}}{N_m!}[(\partial_t^{N_m}F)(x-y,t) - (\partial_t^{N_m}F)(x,t)] \\ &+\frac{(-s)^{N_m+1}}{N_m!}\int_0^1 (1-\tau)^{N_m}(\partial_t^{N_m+1}F)(x-y,t-s\tau)d\tau \end{aligned}$$

if m is odd, and

$$\begin{aligned} S^m = &\sum_{|\beta|+2p=m-1} |\beta| \int_0^1 (1-\theta)^{|\beta|-1} \frac{(-y)^\beta(-s)^p}{p!\beta!} \\ &\qquad\qquad \times[(\partial_x^\beta\partial_t^p F)(x-y\theta,t) - (\partial_x^\beta\partial_t^p F)(x,t)]d\theta \\ &+\frac{(-s)^{N_m+1}}{N_m!}\int_0^1 (1-\tau)^{N_m}(\partial_t^{N_m+1}F)(x-y,t-s\tau)d\tau \end{aligned}$$

if m is even.

The proof of Lemma 5.2 is straightforward, and so it is omitted here.

Proof of Theorem 5.1. We apply Lemma 5.2 to get

$$\begin{aligned} J_1 + &\sum_{|\beta|+2p\leq m-1} \frac{(-1)^{|\beta|+p}}{p!\beta!}(\partial_t^p\partial_x^\beta F_{\ell,jk})(x,t)\int_0^\infty\!\!\int s^p y^\beta (u_\ell u_k)(y,s)dyds \\ &= \sum_{|\beta|+2p\leq m-1} \frac{(-1)^{|\beta|+p}}{p!\beta!}(\partial_t^p\partial_x^\beta F_{\ell,jk})(x,t)\int_{t/2}^\infty\!\!\int s^p y^\beta (u_\ell u_k)(y,s)dyds \\ &\qquad - \int_0^{t/2}\!\!\int S^m_{\ell,jk}(x,y,t,s)(u_\ell u_k)(y,s)dyds \\ &\equiv J_{11} + J_{12}. \end{aligned}$$

Suppose u is a strong solution and recall (2.1) and (2.7), i.e., that

$$\|(\partial_t^p\partial_x^\beta F_{\ell,jk})(\cdot,t)\|_q \leq C_q t^{-\frac{1+|\beta|+2p}{2}-\frac{n}{2}(1-\frac{1}{q})},$$

$$\int s^p|y|^{|\beta|}|u(y,s)|^2dy \leq C(1+s)^{-\frac{n-(|\beta|+2p)}{2}-1}.$$

If $|\beta| + 2p \le m-1$, each term of J_{11} behaves in L^q like $t^{-\frac{n+1}{2}-\frac{n}{2}(1-\frac{1}{q})}$ as $t \to \infty$; hence

$$t^{\frac{m}{2}+\frac{n}{2}(1-\frac{1}{q})}\|J_{11}\|_q \le Ct^{-\frac{n-m+1}{2}} \le Ct^{-\frac{1}{2}} \to 0 \qquad \text{as } t \to \infty$$

for $1 \le q \le \infty$ and $1 \le m \le n$. So we need only estimate J_{12}, using the concrete expression of functions $S^m_{\ell,jk}$ as given in Lemma 5.2. Let S^m_2 be the last term in the definition of S^m which involves the integral in τ, and write

$$S^m = S^m_1 + S^m_2.$$

Since $0 \le s \le t/2$, straightforward estimation shows that

$$\|(\partial_t^{N_m+1}F)(\cdot - y, t - s\tau)\|_q \le ct^{-\frac{3}{2}-N_m-\frac{n}{2}(1-\frac{1}{q})} = \begin{cases} ct^{-1-\frac{m}{2}-\frac{n}{2}(1-\frac{1}{q})} & (m : \text{odd}), \\ ct^{-\frac{1}{2}-\frac{m}{2}-\frac{n}{2}(1-\frac{1}{q})} & (m : \text{even}), \end{cases}$$

and

$$\begin{aligned} \int_0^{t/2} s^{N_m+1}\|u(s)\|_2^2 ds &\le c\int_0^{t/2} s^{N_m+1}(1+s)^{-1-\frac{n}{2}}ds \\ &\le \begin{cases} c\displaystyle\int_0^{t/2}(1+s)^{-\frac{n-m+1}{2}}ds & (m : \text{odd}), \\ c\displaystyle\int_0^{t/2}(1+s)^{-1-\frac{n-m}{2}}ds & (m : \text{even}). \end{cases} \end{aligned}$$

Therefore, the contribution from S^m_2 is estimated as

$$\le Ct^{-\frac{m}{2}-\frac{n}{2}(1-\frac{1}{q})} \times t^{-1}\int_0^t (1+s)^{-\frac{1}{2}}ds = o(t^{-\frac{m}{2}-\frac{n}{2}(1-\frac{1}{q})}) \qquad \text{if } m \text{ is odd,}$$

and

$$\le Ct^{-\frac{m}{2}-\frac{n}{2}(1-\frac{1}{q})} \times t^{-\frac{1}{2}}\int_0^t (1+s)^{-1}ds = o(t^{-\frac{m}{2}-\frac{n}{2}(1-\frac{1}{q})}) \qquad \text{if } m \text{ is even.}$$

To estimate the contribution from S^m_1, we write

$$(\partial_x^\beta \partial_t^p F)(x,t) = t^{-\frac{1+n+|\beta|+2p}{2}}K(xt^{-\frac{1}{2}}),$$

and, when m is odd,

$$(\partial_t^{N_m}F)(x,t) = t^{-\frac{m+n}{2}}K(xt^{-\frac{1}{2}})$$

in terms of some functions K which are smooth, bounded, integrable, and uniformly continuous on $\mathbb{R}^n$. We easily see that

$$\|(\partial_x^\beta \partial_t^p F)(\cdot - y\theta, t) - (\partial_x^\beta \partial_t^p F)(\cdot, t)\|_q \le Ct^{-\frac{m}{2}-\frac{n}{2}(1-\frac{1}{q})}\|K(\cdot - yt^{-\frac{1}{2}}\theta) - K(\cdot)\|_q$$

when $|\beta| + 2p = m-1$, and

$$\|(\partial_t^{N_m}F)(\cdot - y, t) - (\partial_t^{N_m}F)(\cdot, t)\|_q \le Ct^{-\frac{m}{2}-\frac{n}{2}(1-\frac{1}{q})}\|K(\cdot - yt^{-\frac{1}{2}}) - K(\cdot)\|_q.$$

Using these estimates as well as (2.7), we can proceed in exactly the same way as in section 4 to conclude that the contribution from S_1^m is $o(t^{-\frac{m}{2}-\frac{n}{2}(1-\frac{1}{q})})$. We have thus deduced

$$\lim_{t\to\infty} t^{\frac{m}{2}+\frac{n}{2}(1-\frac{1}{q})}\|J_1\|_q = 0 \tag{5.7}$$

for $1 \le m \le n$ and $1 \le q \le \infty$ when u is a strong solution.

When u is a weak solution, we can estimate J_{11} and J_{12} in the same way as above, using (2.8) instead of (2.7), and conclude that (5.7) holds for $1 \le m \le n-1$ and $1 \le q \le \infty$. This completes the proof of Theorem 5.1. □

Remarks.

(i) In this section we could treat higher-order expansions of weak solutions only in space dimensions $n = 3, 4$ because the moment estimate (2.8) is known only in this case. It should be noticed that in Theorem 2.1(ii), we could treat weak solutions in general space dimensions $n \ge 2$ because we then needed (2.8) only with $m = 0$ which is valid in all space dimensions. As will be seen from the argument in the appendix, it seems impossible to deduce (2.8) for $m \ge 1$ when $n \ge 5$. Indeed, the desired boundedness is first deduced for approximate solutions and then for the weak solutions by passing to the limit. However, if the boundedness were true for the approximate solutions on $\mathbb{R}^n$, $n \ge 5$, we could then deduce the precompactness of the approximate solutions in $L^2(0, T : \boldsymbol{L}^2)$ for any fixed $T > 0$. This precompactness readily implies that the weak solutions obtained by passing to the limit satisfy the so-called *strong energy inequality* of Leray [6]:

$$\|u(t)\|_2^2 + 2\int_s^t \|\nabla u\|_2^2 d\tau \le \|u(s)\|_2^2 \qquad \text{for } s = 0, \text{ a.e. } s > 0, \text{ and all } t \ge s.$$

However, the existence of weak solutions satisfying this inequality remains open when $n \ge 5$ and seems in general not to be valid, as is remarked in [5].

(ii) From the argument in this section, we see that if our weak solutions should satisfy the moment estimate (2.7) under suitable conditions on the initial data a, then we could deduce the asymptotic expansion (2.4) with $1 \le q \le 2$ and $1 \le m \le n$ also for the weak solutions. Indeed, we needed moment estimates, which are different between the cases of weak and strong solutions, only in dealing with the integral J_1, and our estimates for J_2 (given in section 4) are independent of the moment estimates.

6. Proof of Theorem 2.3. In this section we prove Theorem 2.3. The case $2 \le q \le \infty$ is treated in [12, 15] by an elementary method. Our main purpose is to extend the result to the case $1 \le q < 2$ by employing the Hardy space theory. Consider the solution ω of the linear heat equation

$$\partial_t \omega = \Delta\omega, \qquad \omega(0) = \omega_0,$$

where $\omega_0(y) = \omega_0(|y|) \in \mathcal{S}(\mathbb{R}^2)$, $\widehat{\omega_0} \in C_c^\infty(\mathbb{R}^2)$, and $\widehat{\omega_0} \equiv 0$ in a neighborhood of $\xi = 0$. Note that $\widehat{\omega_0}$ is also radial, and so both ω_0 and $\widehat{\omega_0}$ can be chosen as real-valued functions. Thus, $\widehat{\omega}(\xi, t) = e^{-t|\xi|^2}\widehat{\omega_0}(\xi)$ is radial, and so $\omega(x, t)$ is also radial. Moreover, $\omega(t) \in \mathcal{S}(\mathbb{R}^2)$ and

$$\int x^\alpha \omega(x, t)dx = \int y^\alpha \omega_0(y)dy = 0 \qquad \text{for all multi-indices } \alpha.$$

This implies that $\omega(t)$ belongs to the Hardy space $\mathcal{H}^p$ for all $0 < p < \infty$ and all $t \geq 0$ ([16, p. 128]), and

$$\|\omega(t)\|_{H^q} \leq Ct^{-(\frac{1}{p}-\frac{1}{q})}\|\omega_0\|_{H^p} \qquad \text{whenever } 0 < p \leq q < \infty. \tag{6.1}$$

See [7] for a proof of (6.1). Consider now

$$u(x,t) = (u_1(x,t), u_2(x,t)) = \frac{1}{2\pi}\int \frac{(-x_2+y_2, x_1-y_1)}{|x-y|^2}\omega(y,t)dy,$$

where $x = (x_1, x_2)$ and $y = (y_1, y_2)$. Then

$$\nabla \times u \equiv \partial_1 u_2 - \partial_2 u_1 = \omega \qquad \text{and} \qquad u \cdot \nabla\omega = 0,$$

so $\partial_t\omega - \Delta\omega + u \cdot \nabla\omega = 0$, and therefore,

$$\nabla \times (\partial_t u - \Delta u + u \cdot \nabla u) = 0.$$

Thus, there exists a scalar function p such that

$$\partial_t u - \Delta u + u \cdot \nabla u + \nabla p = 0.$$

Since $\nabla \cdot u = 0$, it follows that u solves the Navier–Stokes system. Applying the Fourier transform gives

$$\widehat{u}(\xi,t) = \frac{(-i\xi_2, i\xi_1)}{|\xi|^2}\widehat{\omega}(\xi,t) = \frac{(-i\xi_2, i\xi_1)}{|\xi|^2}e^{-t|\xi|^2}\widehat{\omega_0}(\xi).$$

This shows that for each fixed t, the function $\widehat{u}(\xi,t)$ is in $C_c^\infty(\mathbb{R}^2)$ and vanishes in a fixed neighborhood of $\xi = 0$ independent of t. The Hausdorff–Young inequality for the Fourier transform shows that if $2 \leq q \leq \infty$, then

$$\|u(t)\|_q \leq C_q\|\widehat{u}(t)\|_{q'} \leq C_q\left(\int e^{-q't|\xi|^2}|\widehat{\omega_0}(\xi)|^{q'}d\xi\right)^{1-1/q} \leq C_q e^{-\gamma_q t},$$

with $1/q' = 1 - 1/q$, since $\widehat{\omega_0} \equiv 0$ in a neighborhood of $\xi = 0$. We thus conclude that

$$\|u(t)\|_q \leq C_q e^{-\gamma_q t} \qquad \text{for all } 2 \leq q \leq \infty. \tag{6.2}$$

We next write $\widehat{u}$ in the form $\widehat{u}(\xi,t) = |\xi|^{-1}(-i\xi_2|\xi|^{-1}, i\xi_1|\xi|^{-1})\widehat{\omega}(\xi,t)$ so that

$$u = (-\Delta)^{-1/2}(-R_2, R_1)\omega,$$

where R_j are the Riesz transforms. Since R_j are bounded in Hardy spaces, it follows by the Hardy–Littlewood–Sobolev inequality in Hardy spaces [16, p. 136] and (6.1) that $\|u(t)\|_{H^{6/7}} \leq C\|\omega(t)\|_{H^{3/5}} \leq C\|\omega_0\|_{H^{3/5}}$, and therefore

$$\|u(t)\|_1 \leq C\|u(t)\|_{H^1} \leq C\|u(t)\|_2^{1/4}\|u(t)\|_{H^{6/7}}^{3/4} \leq C_1 e^{-\gamma_1 t}.$$

This, together with (6.2), implies

$$\|u(t)\|_q \leq C_q e^{-\gamma t} \qquad \text{for all } 1 \leq q \leq \infty \tag{6.3}$$

with another constant $\gamma > 0$ independent of q. To see the behavior with respect to the space variables, we fix a function $M = (M_1, M_2)$ so that $\widehat{M} \in C_c^\infty(\mathbb{R}^2)$, $\widehat{M} \equiv 0$, in a neighborhood of $\xi = 0$, and

$$\widehat{M}(\xi) = \frac{(-i\xi_2, i\xi_1)}{|\xi|^2} \qquad \text{in a neighborhood of supp}\, \widehat{\omega_0}.$$

Since $\widehat{u}(\xi, t) = \widehat{M}(\xi)e^{-t|\xi|^2}\widehat{\omega_0}(\xi)$, the relation

$$x^\beta u(x,t) = (2\pi)^{-2} \int e^{ix\cdot\xi}(i\partial_\xi)^\beta[\widehat{M}(\xi)e^{-t|\xi|^2}\widehat{\omega_0}(\xi)]d\xi$$

and the fact that $\widehat{\omega_0} \equiv 0$ in a neighborhood of $\xi = 0$ together imply

$$|x^\beta u(x,t)| \le C_\beta \sum_{\eta\le\beta} \int e^{-t|\xi|^2}|\partial_\xi^\eta \widehat{\omega_0}(\xi)|d\xi \le C_\beta e^{-\gamma t} \sum_{\eta\le\beta} \int |\partial_\xi^\eta \widehat{\omega_0}(\xi)|d\xi \le C_\beta e^{-\gamma t}$$

for all multi-indices β. Hence

$$|u(x,t)| \le C_m e^{-\gamma t}(1 + |x|)^{-m} \qquad \text{for all integers } m \ge 0. \tag{6.4}$$

By (6.3) and (6.4) the proof of Theorem 2.3 is complete. □

Appendix. On boundedness of L^2-moments of weak solutions. We shall prove the following, which was employed in the proof of Theorem 2.2(iv).

PROPOSITION A.1. *Let $n = 3$ or 4 and suppose that*

$$\int (1 + |x|)|a(x)|dx < \infty, \qquad \int (1 + |x|)^n |a(x)|^2 dx < \infty.$$

Then the corresponding weak solution u obtained via the methods of [3, 4] *satisfies*

$$\int |x|^n |u(x,t)|^2 dx \le C \qquad \text{for all } \ t \ge 0.$$

Consequently,

$$\int |x|^m |u(x,t)|^2 dx \le C(1+t)^{-(1+\frac{n}{2})(1-\frac{m}{n})} \qquad (m = 0, 1, \dots, n). \tag{A.1}$$

The above result is due to [3, 14]. We here give a detailed proof, modifying slightly the argument of [3], since [3] and [14] are not yet published.

Proof of Proposition A.1. Note that (see [18]) (A.1) is known for $m = 0$. The Navier–Stokes system is

$$\begin{aligned} \partial_t u - \Delta u + u\cdot\nabla u + \nabla p = 0,\\ \nabla\cdot u = 0. \end{aligned} \tag{NS}$$

Assuming, as we may (see [4]), that u is smooth, we multiply (NS) by $2|x|u$ and integrate by parts to get

$$\partial_t \int |x||u|^2 dx + 2\int |x||\nabla u|^2 dx = -2\int \nabla u \cdot \frac{x}{|x|} u dx - 2\int u\cdot\nabla u |x| u dx$$
$$+2\int p\frac{x}{|x|} u dx.$$

Direct calculation gives

$$\int u\cdot\nabla u|x|udx = -\int u|x|u\cdot\nabla udx - \int uu\cdot\frac{x}{|x|}udx$$

so that

$$\int u\cdot\nabla u|x|udx = -\frac{1}{2}\int uu\cdot\frac{x}{|x|}udx.$$

Therefore,

$$\text{(A.2)}\qquad \left|\int u\cdot\nabla u|x|udx\right| \le \frac{1}{2}\|u\|_3^3 \le \begin{cases} C\|u\|_2^{3/2}\|\nabla u\|_2^{3/2} & (n=3),\\ C\|u\|_2\|\nabla u\|_2^2 & (n=4).\end{cases}$$

Furthermore, since

$$\text{(A.3)}\qquad p = R_jR_k(u_ju_k)$$

with $R=(R_1,\dots,R_n)$ the Riesz transforms (see [16]), applying the L^p-boundedness of singular integrals [16] gives

$$\|p\|_2 \le C\|u\|_4^2 \le \begin{cases} C\|u\|_2^{1/2}\|\nabla u\|_2^{3/2} & (n=3),\\ C\|\nabla u\|_2^2 & (n=4).\end{cases}$$

It follows that

$$\text{(A.4)}\qquad \left|\int p\frac{x}{|x|}udx\right| \le \|p\|_2\|u\|_2 \le \begin{cases} C\|u\|_2^{3/2}\|\nabla u\|_2^{3/2} & (n=3),\\ C\|u\|_2\|\nabla u\|_2^2 & (n=4).\end{cases}$$

Finally,

$$\left|\int u\frac{x}{|x|}\nabla udx\right| \le \|u\|_2\|\nabla u\|_2 \le C(\|u\|_2^2+\|\nabla u\|_2^2).$$

Since $\|u\|_2^{3/2}\|\nabla u\|_2^{3/2} \le C(\|u\|_2^6+\|\nabla u\|_2^2) \le C(\|u\|_2^2+\|\nabla u\|_2^2)$, we see from (A.2) and (A.4) that

$$\partial_t\int|x||u|^2dx + \int|x||\nabla u|^2dx \le C(\|u\|_2^2+\|\nabla u\|_2^2)$$

and the right-hand side is integrable in $t\in[0,\infty)$ by (1.6) and (1.7). Hence we get

$$\text{(A.5)}\qquad \int|x||u(x,t)|^2dx \le C.$$

We next multiply (NS) by $2|x|^2|u|$ and integrate by parts to get

$$\partial_t\int|x|^2|u|^2dx + 2\int|x|^2|\nabla u|^2dx = -4\int\nabla u\cdot x\cdot udx - 2\int u\cdot\nabla u|x|^2udx$$
$$+4\int px\cdot udx.$$

We have

$$\left|\int \nabla u\cdot x\cdot udx\right| \le \left(\int |x|^2|\nabla u|^2dx\right)^{1/2}\left(\int |u|^2dx\right)^{1/2}$$
$$\le \varepsilon \int |x|^2|\nabla u|^2dx + C_\varepsilon \int |u|^2dx$$

for all $\varepsilon > 0$. Furthermore,

$$\int u\cdot\nabla u|x|^2udx = -\int uu\cdot x\cdot udx$$

and so

$$\left|\int u\cdot\nabla u|x|^2udx\right| \le \int |x||u|^3dx \le \left(\int |x|^2|u|^2dx\right)^{1/2}\|u\|_4^2$$
$$\le \begin{cases} C\left(\int |x|^2|u|^2dx\right)^{1/2}(\|u\|_2^2+\|\nabla u\|_2^2) & (n=3),\\ C\left(\int |x|^2|u|^2dx\right)^{1/2}\|\nabla u\|_2^2 & (n=4).\end{cases}$$

Similarly,

$$\left|\int px\cdot udx\right| \le \|p\|_2\left(\int |x|^2|u|^2dx\right)^{1/2} \le C\|u\|_4^2\left(\int |x|^2|u|^2dx\right)^{1/2}$$
$$\le \begin{cases} C\left(\int |x|^2|u|^2dx\right)^{1/2}(\|u\|_2^2+\|\nabla u\|_2^2) & (n=3),\\ C\left(\int |x|^2|u|^2dx\right)^{1/2}\|\nabla u\|_2^2 & (n=4).\end{cases}$$

We thus obtain

$$\partial_t\int |x|^2|u|^2dx + \int |x|^2|\nabla u|^2dx \le C\|u\|_2^2 + C(\|u\|_2^2+\|\nabla u\|_2^2)$$
$$+ C(\|u\|_2^2+\|\nabla u\|_2^2)\int |x|^2|u|^2dx.$$

Since $\|u\|_2^2+\|\nabla u\|_2^2$ is integrable in $t\in[0,\infty)$ by (1.6) and (1.7), we get

$$\int |x|^2|u(x,t)|^2dx \le C \tag{A.6}$$

by Gronwall's lemma. Consequently,

$$\int |x||u(x,t)|^2dx \le C(1+t)^{-\frac{1}{2}-\frac{n}{4}}. \tag{A.7}$$

We next multiply (NS) by $2|x|^3u$ and integrate by parts to get

$$\partial_t\int |x|^3|u|^2dx + 2\int |x|^3|\nabla u|^2dx = -6\int \nabla u|x|x\cdot udx$$
$$-2\int u\cdot\nabla u|x|^3udx + 6\int p|x|x\cdot udx.$$

We easily see that

$$\left|\int \nabla u|x|x\cdot u dx\right| \le \left(\int |x|^3|\nabla u|^2 dx\right)^{1/2}\left(\int |x||u|^2 dx\right)^{1/2}$$
$$\le \varepsilon\int |x|^3|\nabla u|^2dx + C_\varepsilon\int |x||u|^2dx$$

for all $\varepsilon > 0$. Furthermore,

$$2\int u\cdot\nabla u|x|^3 u dx = -3\int uu\cdot|x|x\cdot udx$$

and so by (A.6),

$$\left|\int u\cdot\nabla u|x|^3udx\right| \le \left(\int |x|^2|u|^2dx\right)^{1/2}\left(\int |x|^2|u|^4dx\right)^{1/2} \le C\left(\int |x|^2|u|^4dx\right)^{1/2}.$$

Similarly, from

$$\int |x|^2|p|^2dx \le C\int |x|^2|u|^4dx,$$

which shows the boundedness of operators R_j in weighted L^q-spaces [16, p. 218], we see by (A.6) that

$$\left|\int p|x|x\cdot udx\right| \le \left(\int |x|^2|p|^2dx\right)^{1/2}\left(\int |x|^2|u|^2dx\right)^{1/2} \le C\left(\int |x|^2|u|^4dx\right)^{1/2}.$$

When $n = 3$,

$$\begin{aligned}\||x|uu\|_2 &\le \||x|u\|_{24/5}\|u\|_{24/7} \le \|u\|_{24/7}\||x|^{3/2}u\|_6^{2/3}\|u\|_{24/7}^{1/3}\\ &\le C\|u\|_{24/7}^{4/3}\|\nabla(|x|^{3/2}u)\|_2^{2/3}\\ &\le C\|u\|_{24/7}^{4/3}(\||x|^{1/2}u\|_2 + \||x|^{3/2}\nabla u\|_2)^{2/3}\\ &\le C\|u\|_2^{1/2}\|\nabla u\|_2^{5/6}(\||x|^{1/2}u\|_2 + \||x|^{3/2}\nabla u\|_2)^{2/3}\\ &\le C\|u\|_2^{1/2}\|\nabla u\|_2^{5/6}\||x|^{1/2}u\|_2^{2/3} + C\|u\|_2^{1/2}\|\nabla u\|_2^{5/6}\||x|^{3/2}\nabla u\|_2^{2/3}\\ &\le \varepsilon\||x|^{3/2}\nabla u\|_2^2 + C_\varepsilon(\||x|^{1/2}u\|_2^2 + \|u\|_2^2 + \|\nabla u\|_2^2)\end{aligned}$$

for all $\varepsilon > 0$. When $n = 4$, we get

$$\begin{aligned}\||x|uu\|_2 &\le \||x|^{3/2}u\|_4^{2/3}\|u\|_4^{4/3} \le C\|\nabla(|x|^{3/2}u)\|_2^{2/3}\|\nabla u\|_2^{4/3}\\ &\le C(\||x|^{1/2}u\|_2^{2/3} + \||x|^{3/2}\nabla u\|_2^{2/3})\|\nabla u\|_2^{4/3}\\ &\le C_\varepsilon(\|\nabla u\|_2^2 + \||x|^{1/2}u\|_2^2) + \varepsilon\||x|^{3/2}\nabla u\|_2^2\end{aligned}$$

for all $\varepsilon > 0$. Therefore,

$$\partial_t\int |x|^3|u|^2dx + \int |x|^3|\nabla u|^2dx \le C(\||x|^{1/2}u\|_2^2 + \|u\|_2^2 + \|\nabla u\|_2^2).$$

By (1.6), (1.7), and (A.7) the right-hand side is integrable in $t \in [0, \infty)$, and we obtain

$$\int |x|^3 |u(x,t)|^2 dx \le C. \tag{A.8}$$

Consequently,

$$\int |x|^j |u(x,t)|^2 dx \le C(1+t)^{-(1+\frac{n}{2})(1-\frac{j}{3})} \qquad (j = 0, 1, 2, 3, \quad n = 3, 4). \tag{A.9}$$

Assume finally that $n = 4$, multiply (NS) by $2|x|^4 u$, and integrate by parts to get

$$\partial_t \int |x|^4 |u|^2 dx + 2\int |x|^4 |\nabla u|^2 dx = -8 \int \nabla u |x|^2 \cdot x \cdot u dx - 2\int u \cdot \nabla u |x|^4 u dx$$
$$+8 \int p|x|^2 x \cdot u dx.$$

We have

$$\left|\int \nabla u |x|^2 \cdot x \cdot u dx\right| \le \left(\int |x|^4 |\nabla u|^2 dx\right)^{1/2} \left(\int |x|^2 |u|^2 dx\right)^{1/2}$$
$$\le \varepsilon \int |x|^4 |\nabla u|^2 dx + C_\varepsilon \int |x|^2 |u|^2 dx$$

for all $\varepsilon > 0$. Furthermore,

$$2\int u \cdot \nabla u |x|^4 u dx = -4 \int uu|x|^2 x \cdot u dx$$

so that by (A.8)

$$\left|\int u \cdot \nabla u |x|^4 u dx\right| \le C \left(\int |x|^3 |u|^4 dx\right)^{1/2} \left(\int |x|^3 |u|^2 dx\right)^{1/2}$$
$$\le C \left(\int |x|^3 |u|^4 dx\right)^{1/2}.$$

Similarly, from

$$\int |x|^3 |p|^2 dx \le C \int |x|^3 |u|^4 dx,$$

it follows by (A.8) that

$$\left|\int p|x|^2 x \cdot u dx\right| \le C \left(\int |x|^3 |u|^4 dx\right)^{1/2} \left(\int |x|^3 |u|^2 dx\right)^{1/2} \le C \left(\int |x|^3 |u|^4 dx\right)^{1/2}.$$

However,

$$\left(\int |x|^3 |u|^4 dx\right)^{1/2} \le \||x|^2 u\|_4^{3/4} \|u\|_4^{5/4} \le C \|\nabla(|x|^2 u)\|_2^{3/4} \|\nabla u\|_2^{5/4}$$
$$\le C(\||x|u\|_2^{3/4} + \||x|^2 \nabla u\|_2^{3/4}) \|\nabla u\|_2^{5/4}$$
$$\le \varepsilon \||x|^2 \nabla u\|_2^2 + C_\varepsilon (\|\nabla u\|_2^2 + \||x|u\|_2^2)$$

for all $\varepsilon > 0$. We thus have

$$\partial_t \int |x|^4|u|^2 dx + \int |x|^4|\nabla u|^2 dx \le C(\||x|u\|_2^2 + \|\nabla u\|_2^2). \tag{A.10}$$

Since $\||x|u\|_2^2 \le C(1+t)^{-1}$ by (A.9), it follows from (A.10) that

$$\int |x|^4|u(x,t)|^2 dx \le C \log(1+t).$$

However, this implies

$$\||x|u\|_2^2 \le \left(\int |u(x,t)|^2 dx\right)^{1/2} \left(\int |x|^4|u(x,t)|^2 dx\right)^{1/2} \le C(1+t)^{-3/2}[\log(1+t)]^{1/2},$$

and the right-hand side is integrable in $t \in [0,\infty)$. So we conclude from (A.10) that

$$\int |x|^4|u(x,t)|^2 dx \le C \qquad (n=4).$$

This completes the proof of Proposition A.1. □

Remarks.

(i) The above argument also shows that

$$\int_0^\infty \int |x|^n |\nabla u|^2 dx dt < \infty \qquad (n=3,4).$$

(ii) We have omitted discussing the finiteness of L^2-moments. Actually, this is verified first for approximate solutions and the desired boundedness result is then deduced for weak solutions by passing to the limit. The details are described in [3, 14].

(iii) It is impossible to apply the argument of this section to estimating the moments $\int |x|^{n+k}|u|^2 dx$, $k \ge 1$. The main reason is that the pressure p is represented in (A.3) in terms of singular integrals R_j, and one cannot apply in the present situation the weighted L^q-estimates, which were crucial in the above argument, in such a way that the resulting terms are estimated in the Sobolev space $H^1 = W^{1,2}$. To avoid this difficulty, we need some additional integrability conditions on weak solutions; however, conditions of this kind imply that the solutions must be strong solutions. For the details, we refer the reader to [3, 14].

Acknowledgment. We are grateful to the referee for carefully reading the manuscript and sending us valuable comments on the presentation and suggestions on improvements.

REFERENCES

[1] A. Carpio, *Large-time behavior in incompressible Navier–Stokes equations*, SIAM J. Math. Anal., 27 (1996), pp. 449–475.

[2] Y. Fujigaki and T. Miyakawa, *Asymptotic profiles of nonstationary incompressible Navier–Stokes flows in the half-space*, Methods Appl. Anal., to appear.

[3] C. He and Z. Xin, *On the Decay Properties of Solutions to the Nonstationary Navier–Stokes Equations in R^3*, preprint, The Institute of Mathematical Sciences, The Chinese University of Hong Kong, Hong Kong, 1999.

[4] R. Kajikiya and T. Miyakawa, *On L^2 decay of weak solutions of the Navier–Stokes equations in R^n*, Math. Z., 192 (1986), pp. 135–148.

[5] T. Kato, *Strong L^p-solutions of the Navier–Stokes equation in R^m, with applications to weak solutions*, Math. Z., 187 (1984), pp. 471–480.

[6] J. Leray, *Sur le mouvement d'un liquide visqueux emplissant l'éspace*, Acta Math., 63 (1934), pp. 193–248.

[7] T. Miyakawa, *Hardy spaces of solenoidal vector fields, with applications to the Navier–Stokes equations*, Kyushu J. Math., 50 (1996), pp. 1–64.

[8] T. Miyakawa, *Application of Hardy space techniques to the time-decay problem for incompressible Navier–Stokes flows in R^n*, Funkcial. Ekvac., 41 (1998), pp. 383–434.

[9] T. Miyakawa, *On space-time decay properties of nonstationary incompressible Navier–Stokes flows in R^n*, Funkcial. Ekvac., 43 (2000), pp. 541–557.

[10] T. Miyakawa and M. E. Schonbek, *On optimal decay rates for weak solutions to the Navier–Stokes equations in R^n*, Mathematica Bohemica, to appear.

[11] M. E. Schonbek, *Large time behaviour of solutions to the Navier–Stokes equations*, Comm. Partial Differential Equations, 11 (1986), pp. 733–763.

[12] M. E. Schonbek, *Lower bounds of rates of decay for solutions to the Navier–Stokes equations*, J. Amer. Math. Soc., 4 (1991), pp. 423–449.

[13] M. E. Schonbek, *On decay of solutions to the Navier–Stokes equations*, in Applied Nonlinear Analysis, A. Sequeira, H. Beirao da Veiga, and J. H. Videman, eds., Kluwer/Plenum, New York, 1999, pp. 505–512.

[14] M. E. Schonbek and T. P. Schonbek, *On the boundedness and decay of moments of solutions of the Navier–Stokes equations*, Adv. Differential Equations, to appear.

[15] M. E. Schonbek, T. Schonbek, and E. Süli, *Large time behaviour of solutions to the magneto-hydrodynamic equations*, Math. Ann., 304 (1996), pp. 717–756.

[16] E. M. Stein, *Harmonic Analysis*, Princeton University Press, Princeton, NJ, 1993.

[17] S. Takahashi, *A weighted equation approach to decay rate estimates for the Navier–Stokes equations*, Nonlinear Anal., 37 (1999), pp. 751–789.

[18] M. Wiegner, *Decay results for weak solutions of the Navier–Stokes equations in R^n*, J. London Math. Soc. (2), 35 (1987), pp. 303–313.

SIAM J. MATH. ANAL.
Vol. 33, No. 3, pp. 545–569

ON FULLY NONLINEAR PDEs DERIVED FROM VARIATIONAL PROBLEMS OF L^p NORMS[*]

TOSHIHIRO ISHIBASHI[†] AND SHIGEAKI KOIKE[†]

In memory of Professor Yoshihito Tomita

Abstract. The p-Laplace operator arises in the Euler–Lagrange equation associated with a minimizing problem which contains the L^p norm of the gradient of functions. However, when we adapt a different L^p norm equivalent to the standard one in the minimizing problem, a different p-Laplace-type operator appears in the corresponding Euler–Lagrange equation. First, we derive the limit PDE which the limit function of minimizers of those, as $p \to \infty$, satisfies in the viscosity sense. Then we investigate the uniqueness and existence of viscosity solutions of the limit PDE.

Key words. viscosity solution, fully nonlinear equation, ∞-Laplacian, comparison principle, concave solution

AMS subject classifications. 49L25, 35J70, 35J60, 35J20

PII. S0036141000380000

1. Introduction.

We consider the following variational problem:

$$\inf\left\{ \|Dv\|^p_{L^p(\Omega)} - \int_\Omega fv dx \;\middle|\; v \in W^{1,p}_0(\Omega) \right\}, \tag{1.1}$$

where $p > 1$, $\Omega \subset \mathbf{R}^n$ is a bounded domain with smooth boundary $\partial\Omega$ for simplicity, $f : \Omega \to \mathbf{R}$ is a given (smooth) function, Du denotes the gradient of u (i.e., $(u_{x_1}, \ldots, u_{x_n})$), and $\|Du\|^p_{L^p(\Omega)} = \int_\Omega |Du|^p dx$. Here and later, for $\xi = (\xi_1, \ldots, \xi_n) \in \mathbf{R}^n$, we denote $|\xi|$ by the Euclidean norm of ξ; $|\xi|^2 = \sum_{k\in I} \xi_k^2$, where $I = \{1, 2, \ldots, n\}$.

It is well known that the minimizer $u^p \in W^{1,p}_0(\Omega)$ of the variational problem (1.1) is a weak solution (in the distribution sense) of

$$-\sum_{k\in I} p \left(|Du|^{p-2} u_{x_k}\right)_{x_k} = f \quad \text{in } \Omega.$$

In order to deal with a perfect plastic torsion model, it is important to study the limit function of u^p, as $p \to \infty$, when $f \equiv 1$. In fact, in this case, Kawohl in [12] showed that

$$\lim_{p\to\infty} u^p(x) = \operatorname{dist}(x, \partial\Omega) \quad \text{uniformly in } \Omega.$$

We also refer to Bhattacharya, DiBenedetto, and Manfredi [3].

On the other hand, initiated by Aronsson's works [1] and [2], for a given (Lipschitz continuous) function $g : \overline{\Omega} \to \mathbf{R}$, Jensen in [9] characterized the limit function $u := \lim_{p\to\infty} u^p$, where u^p is a minimizer of the variational problem

$$\inf\left\{ \|Dv\|^p_{L^p(\Omega)} \;\middle|\; v - g \in W^{1,p}_0(\Omega) \right\},$$

[*] Received by the editors October 23, 2000; accepted for publication (in revised form) April 11, 2001; published electronically September 5, 2001.
http://www.siam.org/journals/sima/33-3/38000.html

[†] Department of Mathematics, Saitama University, Urawa, Saitama 338-8570, Japan (tishibas@rimath.saitama-u.ac.jp, skoike@rimath.saitama-u.ac.jp). The second author was supported in part by Grant-in-Aid for Scientific Research (12640103 and 12440044) of the Japan Society for the Promotion of Science.

as a unique viscosity solution of

$$\begin{cases} -\triangle_\infty u = 0 & \text{in } \Omega, \\ u = g & \text{on } \partial\Omega, \end{cases} \tag{1.2}$$

where

$$\triangle_\infty u = \sum_{k,l\in I} u_{x_k} u_{x_l} u_{x_k x_l}.$$

In fact, to show the uniqueness of viscosity solutions of (1.2), Jensen introduced variational problems (1.1) with $f = \pm\varepsilon^{p-1}$ (for $\varepsilon > 0$) but under the Dirichlet condition $u = g$ on $\partial\Omega$. Following his argument, we can verify that $\hat{u} := \lim_{p\to\infty} \hat{u}^p \in C(\overline{\Omega})$, where $\hat{u}^p$ is a minimizer of (1.1) when $f \equiv 1$, is a (unique) viscosity solution of

$$\begin{cases} \min\{|Du| - 1, -\triangle_\infty u\} = 0 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

We refer to [3], [5], [6], [10], [11], and [15] for related topics on the L^∞-Laplace equation (1.2).

In this paper, we are interested in the following questions.

(i) If we take other equivalent norms in (1.1) to $\|\cdot\|_{L^p(\Omega)}$, which PDEs are expected to be solved by $\hat{u} := \lim_{p\to\infty} \hat{u}^p$?

(ii) Under which condition can we characterize $\hat{u}$ as a unique (viscosity) solution of the expected PDE?

Let us consider two typical examples.

Example 1.1. If we use the equivalent norm

$$\|Dv\|_{1,p} := \left(\sum_{k\in I} \|v_{x_k}\|^p_{L^p(\Omega)} \right)^{\frac{1}{p}}$$

in (1.1), then we easily verify that the unique minimizer u^p is a weak solution of

$$-p\sum_{k\in I} \left(|u_{x_k}|^{p-2} u_{x_k}\right)_{x_k} = -p(p-1)\sum_{k\in I} |u_{x_k}|^{p-2} u_{x_k x_k} = f \quad \text{in } \Omega. \tag{1.3}$$

In the case when $\min_{\overline{\Omega}} f > 0$ in (1.3), we will observe that $u := \lim_{p\to\infty} u^p$ is a viscosity solution of

$$\min\left\{ \max_{j\in I} |u_{x_j}| - 1, -\sum_{k\in \hat{I}[Du(x)]} u^2_{x_k} u_{x_k x_k} \right\} = 0 \quad \text{in } \Omega, \tag{1.4}$$

where for $\xi = (\xi_1, \ldots, \xi_n) \in \mathbf{R}^n$ we set

$$\hat{I}[\xi] = \left\{ k \in I \;\middle|\; \max_{j\in I} |\xi_j| = |\xi_k| \right\}.$$

We notice that the PDE (1.4) has serious discontinuity with respect to Du-variables.

We note that if we suppose that $\max_{\overline{\Omega}} f < 0$, then the limit function satisfies the following PDE:

$$\max\left\{1 - \max_{j\in I}|u_{x_j}|, -\sum_{k\in\hat{I}[Du(x)]} u_{x_k}^2 u_{x_k x_k}\right\} = 0 \quad \text{in } \Omega.$$

Example 1.2. Next, if we use the equivalent norm

$$\|Dv\|_{\infty,p} := \left(\max_{k\in I}\|v_{x_k}\|_{L^p(\Omega)}^p\right)^{\frac{1}{p}}$$

in (1.1), then it turns out that the minimizer u^p of (1.1) satisfies

$$\max_{k\in I}\int_\Omega \left(p|u_{x_k}|^{p-2}u_{x_k}\phi_{x_k} - f\phi\right)dx \geq 0 \quad \forall\phi\in W_0^{1,p}(\Omega).$$

Hence it holds that

$$\min_{k\in I}\int_\Omega \left(p|u_{x_k}|^{p-2}u_{x_k}\phi_{x_k} - f\phi\right)dx \leq 0 \quad \forall\phi\in W_{0,+}^{1,p}(\Omega) \tag{1.5}$$

and

$$\max_{k\in I}\int_\Omega \left(p|u_{x_k}|^{p-1}u_{x_k}\phi_{x_k} - f\phi\right)dx \geq 0 \quad \forall\phi\in W_{0,+}^{1,p}(\Omega), \tag{1.6}$$

where $W_{0,+}^{1,p}(\Omega) = \{\phi\in W_0^{1,p}(\Omega) \mid \phi\geq 0 \text{ in } \Omega\}$. Thus, if u^p is smooth enough, then it holds that

$$\min_{k\in I}\left\{-p\left(|u_{x_k}|^{p-2}u_{x_k}\right)_{x_k}\right\} = \min_{k\in I}\left\{-p(p-1)|u_{x_k}|^{p-2}u_{x_k x_k}\right\} \leq f \tag{1.7}$$

and

$$\max_{k\in I}\left\{-p\left(|u_{x_k}|^{p-2}u_{x_k}\right)_{x_k}\right\} = \max_{k\in I}\left\{-p(p-1)|u_{x_k}|^{p-2}u_{x_k x_k}\right\} \geq f. \tag{1.8}$$

Therefore, we may call u^p a weak subsolution of (1.7) (resp., a weak supersolution of (1.8)) if u^p satisfies (1.5) (resp., (1.6)).

Let us consider the case when $\min_{\overline{\Omega}} f > 0$ as in Example 1.1.

We will observe that $u := \lim_{p\to\infty} u^p$ is, respectively, a viscosity subsolution and a viscosity supersolution of

$$\min\left\{\max_{j\in I}|u_{x_j}| - 1, F^-(Du, D^2u)\right\} \leq 0 \quad \text{in } \Omega \tag{1.9}$$

and

$$\min\left\{\max_{j\in I}|u_{x_j}| - 1, F^+(Du, D^2u)\right\} \geq 0 \quad \text{in } \Omega, \tag{1.10}$$

where, for $(q, X)\in\mathbf{R}^n\times S^n$,

$$F^-(q,X) = \begin{cases} \min_{k\in I}(-q_k^2 X_{kk}), & \text{provided } \hat{I}[q] = I, \\ \min_{k\in\hat{I}[q]}(-q_k^2 X_{kk})\wedge 0 & \text{otherwise} \end{cases}$$

and

$$F^+(q,X) = \begin{cases} \max\limits_{k\in I}(-q_k^2 X_{kk}), & \text{provided } \hat{I}[q] = I, \\ \max\limits_{k\in \hat{I}[q]}(-q_k^2 X_{kk}) \vee 0 & \text{otherwise.} \end{cases}$$

Here S^n denotes the set of $n \times n$ symmetric matrices equipped with the standard order.

We remark that if $u \in C^2(\Omega)$ satisfies $\hat{I}[Du(x)] \neq I$ for all $x \in \Omega$, then it is a viscosity subsolution of (1.9). Thus we cannot expect the comparison principle for viscosity subsolutions of (1.9) and viscosity supersolutions of (1.10) in general.

We note that in the case when $\min_{\overline{\Omega}} f < 0$, the corresponding limit function satisfies the following inequalities in the viscosity sense:

$$\max\left\{1 - \max_{j\in I}|u_{x_j}|, F^-(Du, D^2u)\right\} \leq 0$$

and

$$\max\left\{1 - \max_{j\in I}|u_{x_j}|, F^+(Du, D^2u)\right\} \geq 0.$$

This paper is organized as follows. In section 2, we recall the definition of viscosity solutions for second-order degenerate elliptic PDEs and its equivalent definitions. We also show that minimizers of variational problems are viscosity solutions of the associated Euler–Lagrange equations.

We verify that the limit function (as $p \to \infty$) of minimizers of our variational problems is indeed a viscosity solution of the corresponding limit PDE in section 3. This verification result immediately gives the existence of viscosity solutions of the limit PDE.

In section 4, we show a comparison result between a viscosity subsolution u and a viscosity supersolution v of the limit PDE, assuming that u or $-v$ is "locally" convex. This comparison principle yields a uniqueness result for locally concave viscosity solutions.

Section 5 is devoted to the existence of concave viscosity solutions when the domain Ω is convex.

In the final section, we study some typical examples.

2. Preliminaries.

2.1. Notations. We shall briefly recall the definition of viscosity solutions of fully nonlinear second-order (degenerate elliptic) PDEs:

$$F(Du, D^2u) = 0 \quad \text{in } \Omega, \tag{2.1}$$

where $F : \mathbf{R}^n \times S^n \to \mathbf{R}$ is given. Although we could deal with more general PDEs (e.g., PDEs having x and $u(x)$ variables), we restrict our attention here to the above PDEs for the sake of simplicity.

We denote by F^* and F_*, respectively, the upper and lower semicontinuous envelopes of F; for $(q, X) \in \mathbf{R}^n \times S^n$,

$$F^*(q,X) = \lim_{\varepsilon\to 0}\sup\{F(q',X') \mid |q'-q| + |X'-X| < \varepsilon\},$$

and $F_*(q,X) = -(-F)^*(q,X)$.

Definition. We call $u \in C(\Omega)$ a viscosity subsolution (resp., supersolution) of (2.1) if

$$F_*(D\phi(x), D^2\phi(x)) \le 0 \quad \big(\text{resp., } F^*(D\phi(x), D^2\phi(x)) \ge 0\big)$$

whenever $u - \phi$ attains its maximum (resp., minimum) at $x \in \Omega$ for $\phi \in C^2(\Omega)$.

We call $u \in C(\Omega)$ a viscosity solution of (2.1) if it is a viscosity subsolution and a viscosity supersolution of (2.1).

Remark. In the above definition, we can change "maximum (resp., minimum)" to "strict maximum (resp., strict minimum)." We refer to [4] for the general theory of viscosity solutions of fully nonlinear second-order degenerate elliptic PDEs.

Since we will use a definition equivalent to the one above, we prepare some notation. For a function $u : \Omega \to \mathbf{R}$ and $x \in \Omega$, we define semijets $J^{2,\pm}u(x)$ in the following manner:

$$J^{2,+}u(x) = \left\{ (q, X) \in \mathbf{R}^n \times S^n \;\middle|\; \begin{array}{ll} u(y) \le & u(x) + \langle q, y - x\rangle \\ & +\langle X(y-x), y-x\rangle/2 \\ & +o(|y-x|^2) \text{ as } y \in \Omega \to x \end{array} \right\},$$

and $J^{2,-}u(x) = -J^{2,+}(-u)(x)$.

For $x \in \Omega$, we denote by $\overline{J}^{2,\pm}u(x)$ the following sets:

$$\left\{ (q, X) \in \mathbf{R}^n \times S^n \;\middle|\; \begin{array}{c} \exists x^m \in \Omega \text{ and } \exists (q^m, X^m) \in J^{2,\pm}u(x^m) \text{ such that} \\ \lim\limits_{m\to\infty} (x^m, u(x^m), q^m, X^m) = (x, u(x), q, X) \end{array} \right\}.$$

PROPOSITION 2.1 (see [4]). *A function $u \in C(\Omega)$ is a viscosity subsolution* (*resp., supersolution*) *of* (2.1) *if and only if*

$$F_*(q, X) \le 0 \quad \forall x \in \Omega \textit{ and } \forall (q, X) \in \overline{J}^{2,+}u(x)$$
$$\left(\textit{resp., } F^*(q, X) \ge 0 \quad \forall x \in \Omega \textit{ and } \forall (q, X) \in \overline{J}^{2,-}u(x)\right).$$

When u is a viscosity subsolution (resp., supersolution) of (2.1), we will often call u a viscosity solution of $F(Du, D^2u) \le 0$ (resp., $F(Du, D^2u) \ge 0$).

In what follows, we shall omit the term "viscosity" since we discuss only viscosity solutions.

For later convenience, we introduce some terminology: First, we denote by $B_r(x)$ the standard open ball in $\mathbf{R}^n$ with radius $r > 0$ and center $x \in \mathbf{R}^n$; $B_r(x) = \{y \in \mathbf{R}^n \mid |x - y| < r\}$. We will write B_r for $B_r(0)$.

Definition. We call a function $u : \Omega \to \mathbf{R}$ locally convex (resp., locally concave) if for each $x \in \Omega$, there is $r > 0$ such that u is convex (resp., concave) in $B_r(x) \subset \Omega$.

We remark that if Ω is a convex domain in $\mathbf{R}^n$, then the local convexity is equivalent to the convexity. However, this equivalence does not hold for general domains. In general, even if $u : \Omega \to \mathbf{R}$ is locally convex in a nonconvex domain, we cannot extend u to the convex-hull of Ω by a convex function.

In section 4, we will use the following basic lemma.

LEMMA 2.2. *Assume that $u : \Omega \to \mathbf{R}$ is locally convex. If $(q, X) \in \overline{J}^{2,+}u(x)$ for $x \in \Omega$, then we have $X \ge 0$.*

Proof. It suffices to show that if $(q, X) \in J^{2,+}u(x)$ for $x \in \Omega$, then $X \ge 0$.

Fix any $(q, X) \in J^{2,+}u(x)$ for $x \in \Omega$.

Since for any $\xi \in \mathbf{R}^n$ with $|\xi| = 1$ and small $t > 0$, the definition yields

$$u(x \pm t\xi) \leq u(x) \pm t\langle q, \xi\rangle + \frac{t^2}{2}\langle X\xi, \xi\rangle + o(t^2),$$

the local convexity implies that

$$0 \leq t^2\langle X\xi, \xi\rangle + o(t^2). \qquad \square$$

2.2. Verification for fixed $p > n$. In this subsection, we fix $p > n$.

In order to treat the examples in section 1 at the same time, throughout this paper we shall consider the norm $\|\cdot\|$ of a function $h = (h_1, \ldots, h_n) \in L^p(\Omega; \mathbf{R}^n)$; for fixed sets $I_j \subset I$ $(j = 1, 2, 3)$,

$$\|h\| := \left(\|P_1[h]\|^p_{L^p(\Omega)} + \sum_{k \in I_2} \|h_k\|^p_{L^p(\Omega)} + \max_{k \in I_3} \|h_k\|^p_{L^p(\Omega)}\right)^{\frac{1}{p}}.$$

Here $P_1[h] = \sum_{k \in I_1} h_k e_k$, where e_k is the standard kth unit vector.

We shall suppose that

$$I_1 \cup I_2 \cup I_3 = I := \{1, 2, \ldots, n\}. \tag{A1}$$

Although there are (infinitely) many other equivalent norms, we have decided to choose the above one in this paper since it contains three typical seminorms.

We consider the variational problem

$$\inf\left\{\|Dv\|^p - \int_\Omega fv dx \;\middle|\; v \in W_0^{1,p}(\Omega)\right\}. \tag{2.2}$$

It is well known that if we suppose that $f \in C(\overline{\Omega})$, for instance, then under (A1), there is a (unique) minimizer $u^p \in W_0^{1,p}(\Omega)$ of this variational problem since this norm is equivalent to the standard one, which is $\|\cdot\|_{W^{1,p}(\Omega)}$.

We introduce continuous functions $F_p^\pm : \mathbf{R}^n \times S^n \to \mathbf{R}$ for $p > 1$, which arise in the Euler–Lagrange equation associated with the variational problem (2.2):

$$\begin{aligned} F_p^-(q, X) = \; & -p|P_1[q]|^{p-4} \sum_{k,l \in I_1} \left(\delta_{kl} q_k^2 + (p-2) q_k q_l\right) X_{kl} \\ & + p(p-1)\left(-\sum_{k \in I_2} |q_k|^{p-2} X_{kk} + \min_{k \in I_3}\left(-|q_k|^{p-2} X_{kk}\right)\right) \end{aligned}$$

and

$$\begin{aligned} F_p^+(q, X) = \; & -p|P_1[q]|^{p-4} \sum_{k,l \in I_1} \left(\delta_{kl} q_k^2 + (p-2) q_k q_l\right) X_{kl} \\ & + p(p-1)\left(-\sum_{k \in I_2} |q_k|^{p-2} X_{kk} + \max_{k \in I_3}\left(-|q_k|^{p-2} X_{kk}\right)\right). \end{aligned}$$

PROPOSITION 2.3. *Assume that* (A1) *holds. Then, for any $f \in C(\overline{\Omega})$, the minimizer u^p of the above variational problem* (2.2) *is, respectively, a sub- and a supersolution of*

$$F_p^-(Du, D^2u) - f \leq 0 \tag{2.3}$$

and

$$F_p^+(Du, D^2u) - f \geq 0. \tag{2.4}$$

Remarks. (1) If $I_2 = I_3 = \emptyset$, then $F_p^- = F_p^+$ becomes the standard p-Laplace operator;

$$F_p^\pm(q, X) = -p \sum_{k,l \in I} \left(\delta_{kl}|q|^{p-2} + (p-2)|q|^{p-4} q_k q_l\right) X_{kl}.$$

(2) If $I_1 = I_3 = \emptyset$, then $F_p^- = F_p^+$ becomes the PDE part of (1.3) in Example 1.1.

(3) If $I_1 = I_2 = \emptyset$, then we have that $F_p^\pm$ are the PDE parts of (1.7) and (1.8) in Example 1.2.

Proof. We shall show only the assertion for subsolutions since the proof for supersolutions can be done in a symmetric way.

Suppose that there is $\phi \in C^2(\Omega)$ such that $u^p(\hat{x}) = \phi(\hat{x})$ for some $\hat{x} \in \Omega$, $u^p(x) < \phi(x)$ for $x \in \Omega \setminus \{\hat{x}\}$, and

$$F_p^-(D\phi(\hat{x}), D^2\phi(\hat{x})) \geq f(\hat{x}) + 2\theta \quad \text{for some } \theta > 0.$$

Then we will get a contradiction.

We may suppose that $\hat{x} = 0 \in \Omega$. Let us simply write u for the minimizer u^p.

Choosing smaller $\theta > 0$ if necessary, we can find small $r > t > 0$ such that $B_r \subset \Omega$,

$$F_p^-(D\phi(x), D^2\phi(x)) \geq f(x) + \theta \quad \text{in } B_r,$$
$$\min_{B_t}(u - \phi) \geq -\theta \quad \text{and} \quad \max_{\partial B_r}(u - \phi) \leq -3\theta.$$

Hence, for any $\psi \in W_{0,+}^{1,p}(B_r)$, by integration by parts, we have

$$\begin{aligned} &\sum_{k \in I_1} \int_{B_r} |P_1[D\phi]|^{p-2} \phi_{x_k} \psi_{x_k} dx + \sum_{k \in I_2} \int_{B_r} |\phi_{x_k}|^{p-2} \phi_{x_k} \psi_{x_k} dx \\ &\qquad + \min_{k \in I_3} \int_{B_r} |\phi_{x_k}|^{p-2} \phi_{x_k} \psi_{x_k} dx \geq \frac{1}{p} \int_{B_r} (f + \theta) \psi dx. \end{aligned} \tag{2.5}$$

On the other hand, since u is a minimizer of (2.2), for any $\psi \in W_{0,+}^{1,p}(\Omega)$, we have

$$\begin{aligned} &\sum_{k \in I_1} \int_\Omega |P_1[Du]|^{p-2} u_{x_k} \psi_{x_k} dx + \sum_{k \in I_2} \int_\Omega |u_{x_k}|^{p-2} u_{x_k} \psi_{x_k} dx \\ &\qquad + \min_{k \in I_3} \int_\Omega |u_{x_k}|^{p-2} u_{x_k} \psi_{x_k} dx \leq \frac{1}{p} \int_\Omega f \psi dx. \end{aligned} \tag{2.6}$$

Setting

$$\psi(x) = \begin{cases} (u(x) - \phi(x) + 2\theta)^+ & \text{for } x \in B_r, \\ 0 & \text{for } x \in \Omega \setminus B_r, \end{cases}$$

if $I_3 \neq \emptyset$, by (2.5) and (2.6) we can find $j_0 \in I_3$ such that

$$\begin{aligned} -\theta^2 \int_{B_t} dx \geq &\sum_{k \in I_1} \int_{B_r} \left(|P_1[Du]|^{p-2} u_{x_k} - |P_1[D\phi]|^{p-2}\right) \psi_{x_k} dx \\ &+ \sum_{k \in I_2} \int_{B_r} \left(|u_{x_k}|^{p-2} u_{x_k} - |\phi_{x_k}|^{p-2} \phi_{x_k}\right) \psi_{x_k} dx \\ &+ \int_{B_r} \left(|u_{x_{j_0}}|^{p-2} u_{x_{j_0}} - |\phi_{x_{j_0}}|^{p-2} \phi_{x_{j_0}}\right) \psi_{x_{j_0}} dx, \end{aligned}$$

which is a contradiction.

In the case when $I_3 = \emptyset$, we need only to delete the third term above. □

3. Verification. In this section, we seek the PDEs which, as $p \to \infty$, the limit of minimizers of (2.2) satisfies in the viscosity sense when f is positive. In fact, the "limit" PDEs will be derived from (2.3) and (2.4).

3.1. The limit PDEs. We introduce some notation: for $q = (q_1, \ldots, q_n) \in \mathbf{R}^n$, we set

$$q_0 = |P_1[q]| \quad \text{and} \quad G(q) = \max\{|q_j| \mid j \in \{0\} \cup I_2 \cup I_3\}.$$

Moreover, we set

$$I[q] = \{k \in \{0\} \cup I_2 \cup I_3 \mid G(q) = |q_k|\},$$
$$I_1[q] = \begin{cases} I_1, & \text{provided } G(q) = q_0, \\ \emptyset, & \text{provided } G(q) > q_0, \end{cases}$$

and

$$I_k[q] = \{j \in I_k \mid G(q) = |q_j|\} \quad \text{for } k = 2, 3.$$

For $J \subset I_3$, we set

$$F^-_J(q, X) = \begin{cases} 0, & \text{provided } J = \emptyset, \\ \min\limits_{k \in I_3}(-q_k^2 X_{kk}), & \text{provided } J = I_3, \\ \min\limits_{k \in J}(-q_k^2 X_{kk}) \wedge 0 & \text{otherwise} \end{cases}$$

and

$$F^+_J(q, X) = \begin{cases} 0, & \text{provided } J = \emptyset, \\ \max\limits_{k \in I_3}(-q_k^2 X_{kk}), & \text{provided } J = I_3, \\ \max\limits_{k \in J}(-q_k^2 X_{kk}) \vee 0 & \text{otherwise.} \end{cases}$$

Using this notation, we define (possibly discontinuous) functions $F^\pm_\infty : \mathbf{R}^n \times S^n \to \mathbf{R}$ in the following way: for $(q, X) \in \mathbf{R}^n \times S^n$, we set

$$F^\pm_\infty(q, X) = -\sum_{k,l \in I_1[q]} q_k q_l X_{kl} - \sum_{k \in I_2[q]} q_k^2 X_{kk} + F^\pm_{I_3[q]}(q, X).$$

Here and later, whenever $I_k[q] = \emptyset$ for $k \in \{1, 2\}$, we regard the terms containing $I_k[q]$ as zero.

THEOREM 3.1. *Assume that* (A1) *holds. Assume also that $f \in C(\overline{\Omega})$ satisfies*

$$\min_{x \in \overline{\Omega}} f(x) > 0.$$

Let $u^p \in W^{1,p}_0(\Omega)$ be the minimizer of the variational problem (2.2). *Then there exists a subsequence u^{p_i} such that u^{p_i} converges to a $u \in W^{1,\infty}(\Omega)$ uniformly in $\overline{\Omega}$. Moreover, we see that $u = 0$ on $\partial\Omega$ and u is a solution of*

$$\begin{cases} \min\{G(Du) - 1, F^-_\infty(Du, D^2u)\} \leq 0 & \text{in } \Omega, \\ \min\{G(Du) - 1, F^+_\infty(Du, D^2u)\} \geq 0 & \text{in } \Omega. \end{cases}$$

Before giving a proof of Theorem 3.1, we present representation formulas of $(F_\infty^+)^*$ and $(F_\infty^-)_*$, which will be proved after the proof of Theorem 3.1.

LEMMA 3.2. *$(F_\infty^+)^*(q,X)$ and $(F_\infty^-)_*(q,X)$ are, respectively, represented by*

$$\max\left\{ -\sum_{k,l\in I_1} 1_J q_k q_l X_{kl} - \sum_{k\in I_2\cap J} q_k^2 X_{kk} + F^+_{I_3\cap J}(q,X) \;\middle|\; \emptyset \neq J \subset I[q] \right\}$$

and

$$\min\left\{ -\sum_{k,l\in I_1} 1_J q_k q_l X_{kl} - \sum_{k\in I_2\cap J} q_k^2 X_{kk} + F^-_{I_3\cap J}(q,X) \;\middle|\; \emptyset \neq J \subset I[q] \right\},$$

where

$$1_J = \begin{cases} 1, & \textit{provided } 0 \in J, \\ 0 & \textit{otherwise.} \end{cases}$$

In our proof of Theorem 3.1, we will use the following proposition.

PROPOSITION 3.3. (1) *Assume that $(F_\infty^+)^*(q,X) \le -\theta$ for $\theta > 0$. Then we have the following properties:*

$$\text{(i)} \qquad \textit{If } G(q) = q_0, \textit{ then } -\sum_{k,l\in I_1} q_k q_l X_{kl} \le -\frac{\theta}{2}.$$

$$\text{(ii)} \qquad -q_k^2 X_{kk} \le -\theta \textit{ for } k \in I_2[q] \cup I_3[q].$$

(2) *Assume that $(F_\infty^-)_*(q,X) \ge \theta$ for $\theta > 0$. Then we have the following properties:*

$$\text{(i}'\text{)} \qquad \textit{If } G(q) = q_0, \textit{ then } -\sum_{k,l\in I_1} q_k q_l X_{kl} \ge \frac{\theta}{2}.$$

$$\text{(ii}'\text{)} \qquad -q_k^2 X_{kk} \ge \theta \textit{ for } k \in I_2[q] \cup I_3[q].$$

The proof of Proposition 3.3 will also be given in the end this section.

Proof of Theorem 3.1. First, we note that we may suppose that I_1 consists of at least two integers if it is not empty. In fact, if I_1 has only one element, we can follow the argument below replacing I_1 and I_2, respectively, by $\emptyset$ and $I_1 \cup I_2$. Moreover, we may suppose that $I_1 \neq \emptyset$, $I_2 \neq \emptyset$, and $I_3 \neq \emptyset$ since the argument below becomes easier if some of them are empty.

By Proposition 2.3, for $p > n$, we verify that u^p is a solution of

$$\begin{cases} F_p^-(Du, D^2u) - f \le 0 & \text{in } \Omega, \\ F_p^+(Du, D^2u) - f \ge 0 & \text{in } \Omega. \end{cases}$$

Moreover, since $0 \in W_0^{1,p}(\Omega)$, we find a constant $C > 0$ independent of $p > n$ such that

$$\|Du^p\|_{L^p(\Omega)} \le C\|Du^p\| \le C\|u^p\|_{L^1(\Omega)}^{\frac{1}{p}}.$$

Since $\|u^p\|_{L^1(\Omega)} \le |\Omega|^{(p-1)/p}\|u^p\|_{L^p(\Omega)}$ by the Hölder inequality, the Poincaré inequality (e.g., [7, p. 164]) implies that

$$\|Du^p\|_{L^p(\Omega)} \le C^{\frac{p}{p-1}} |\Omega|^{\frac{1}{p}+\frac{1}{n(p-1)}} |B_1|^{\frac{-1}{n(p-1)}}.$$

Thus, by the Poincaré inequality again, there is a constant $C' > 0$ independent of $p > n$ such that

$$\|u^p\|_{W^{1,p}(\Omega)} \leq C'.$$

Hence, by the Hölder inequality again, we can find a constant $C_0 > 0$ such that

$$\|u^p\|_{W^{1,\hat{p}}(\Omega)} \leq C_0 \quad \forall p \geq \hat{p} > n.$$

Therefore, the Sobolev imbedding implies that there exists $u \in \cap_{p>1} W^{1,p}(\Omega)$ such that u^p converges to u uniformly in $\overline{\Omega}$ by taking a subsequence $\{p^i\}$ such that $\lim_{i\to\infty} p^i = \infty$.

Moreover, from the above inequality, we have $u \in W^{1,\infty}(\Omega)$.

We shall first verify that u is a supersolution of

$$\min\{G(Du) - 1, F^+_\infty(Du, D^2u)\} \geq 0 \quad \text{in } \Omega.$$

Suppose that, for $\phi \in C^2(\Omega)$, $u(\hat{x}) = \phi(\hat{x})$ for $\hat{x} \in \Omega$, $u(x) > \phi(x)$ for $x \in \Omega \setminus \{\hat{x}\}$, and

$$\lim_{\varepsilon\to 0} \sup \left\{ \min\{G(q') - 1, F^+_\infty(q', X')\} \mid |q' - q| + |X' - X| < \varepsilon \right\} \leq -\theta \tag{3.1}$$

for some $\theta > 0$, where $q = D\phi(\hat{x})$ and $X = D^2\phi(\hat{x})$. We will get a contradiction.

In view of the uniform convergence of u^p to u, we may choose $x^p \in \Omega$ such that $u^p - \phi$ attains its minimum at x^p and that $x^p \to \hat{x}$ as $p \to \infty$. Let us simply write (q^p, X^p) for $(D\phi(x^p), D^2\phi(x^p))$.

Consider the case when

$$G(q) < 1.$$

We may suppose that there is $\alpha \in (0,1)$ such that

$$G(q^p) < 1 - \alpha \quad \text{for large } p > 2 \vee n.$$

By Proposition 2.3, we see that

$$\begin{aligned} \frac{f(x^p)}{p(p-2)} \leq & -|q^p_0|^{p-4} \sum_{k,l\in I_1} \left(\frac{\delta_{kl}}{p-2}(q^p_k)^2 + q^p_k q^p_l \right) X^p_{kl} \\ & - \frac{p-1}{p-2} \sum_{k\in I_2} |q^p_k|^{p-2} X^p_{kk} + \frac{p-1}{p-2} \max_{k\in I_3}(-|q^p_k|^{p-2} X^p_{kk}) \\ =: & \, A_1 + A_2 + A_3. \end{aligned} \tag{3.2}$$

From (3.2), we see that $G(q^p) > 0$. Thus, dividing (3.2) by $G(q^p)^{p-4}$, we observe that the left-hand side of the resulting inequality tends to ∞, as $p \to \infty$, while the right-hand side of it is finite.

Thus we may suppose that

$$G(q) \geq 1.$$

Hence, by (3.1), we may suppose that

$$(F^+_\infty)^*(q, X) \leq -\theta. \tag{3.3}$$

Now we shall estimate $\hat{A}_k := A_k/G(q^p)^{p-4}$ $(k=1,2,3)$ from above.

We point out here that

$$\lim_{p\to\infty}\left(\frac{|q^p_k|}{G(q^p)}\right)^{p-4}=0 \quad \text{for } k\in\{0\}\cup I_2\cup I_3, \text{ provided } G(q)>|q_k|. \tag{3.4}$$

Taking a subsequence of $\{q^p\}_{p>n}$ if necessary, we may suppose that there is a nonempty set $\hat{I}\subset\{0\}\cup I_2\cup I_3$ such that $\hat{I}=I[q^p]$ for all $p>n$. We note that $\hat{I}\subset I[q]$ since G is continuous.

Estimate for $\hat{A}_1$. We first remark that for large $p>2\vee n$,

$$\sum_{k\in I_1}\left|\frac{q^p_0}{G(q^p)}\right|^{p-4}\frac{|q^p_k|^2|X_{kk}|}{p-2}\le O(p^{-1}). \tag{3.5}$$

In the case when $0\in\hat{I}$, by (i) of Proposition 3.3, (3.5) implies that for large $p>n$,

$$\hat{A}_1\le-\sum_{k,l\in I_1}q^p_kq^p_lX^p_{kl}+O(p^{-1})=-\sum_{k,l\in I_1}q_kq_lX_{kl}+O(p^{-1})\le-\frac{\theta}{3}.$$

In the case when $0\notin\hat{I}$ and $G(q)=q_0$, since $-\sum_{k,l\in I_1}q^p_kq^p_lX^p_{kl}<0$ by (i) of Proposition 3.3, (3.4) yields that for large $p>n$,

$$\hat{A}_1\le O(p^{-1}).$$

Also, if $G(q)>q_0$, (3.4) yields the same inequality as above.

Therefore, we may suppose that

$$\hat{A}_1\le\begin{cases}-\dfrac{\theta}{3}, & \text{provided } 0\in\hat{I},\\ O(p^{-1}), & \text{provided } 0\notin\hat{I}.\end{cases} \tag{3.6}$$

Estimate for $\hat{A}_2$. We set

$$\begin{aligned}\frac{p-2}{p-1}\hat{A}_2 &= -\sum_{k\in I_2[q]}\left(\frac{|q^p_k|}{G(q^p)}\right)^{p-4}(q^p_k)^2X^p_{kk}-\sum_{k\in I_2\setminus I_2[q]}\left(\frac{|q^p_k|}{G(q^p)}\right)^{p-4}(q^p_k)^2X^p_{kk}\\ &=: C_1+C_2.\end{aligned}$$

In the case when $I_2[q]=\emptyset$ (resp., $I_2\setminus I_2[q]\neq\emptyset$), the above equality holds by taking $C_1=0$ (resp., $C_2=0$).

We note that $-(q^p_k)^2X^p_{kk}<0$ for $k\in I_2[q]$ and large $p>n$ by (ii) of Proposition 3.3.

If there is $k\in\hat{I}\cap I_2$, then we have

$$C_1\le-\sum_{k\in\hat{I}\cap I_2}q_k^2X_{kk}+O(p^{-1})\le-\frac{\theta}{2}.$$

In the case when $\hat{I}\cap I_2=\emptyset$ and $I_2[q]\neq\emptyset$, since $-(q^p_k)^2X^p_{kk}<0$ for $k\in I_2[q]$, we see that for large $p>n$, $C_1\le 0$.

On the other hand, if $I_2\setminus I_2[q]\neq\emptyset$, then (3.4) yields that

$$C_2\le O(p^{-1}).$$

Hence, if $\hat{I}\cap I_2[q]\neq\emptyset$ for large $p>n$, we have

$$\hat{A}_2\leq\frac{p-2}{p-1}\left(-\frac{\theta}{2}+O(p^{-1})\right)\leq-\frac{\theta}{3}.$$

Therefore, we may suppose that

$$\hat{A}_2\leq\begin{cases}-\dfrac{\theta}{3}, & \text{provided } \hat{I}\cap I_2\neq\emptyset,\\ O(p^{-1}), & \text{provided } \hat{I}\cap I_2=\emptyset.\end{cases}\tag{3.7}$$

Estimate for $\hat{A}_3$. If $I_3[q]=\emptyset$, we see that $\hat{A}_3\leq O(p^{-1})$ for large $p>n$.

Hence by (ii) of Proposition 3.3, we see that for large $p>n$,

$$\hat{A}_3\leq\begin{cases}-\dfrac{\theta}{3}, & \text{provided } \hat{I}\cap I_3=I_3,\\ O(p^{-1}), & \text{provided } \hat{I}\cap I_3\neq I_3.\end{cases}\tag{3.8}$$

Let us get a contradiction to (3.3).

First, suppose that $0\in\hat{I}$. Using the estimates (3.6), (3.7), and (3.8) in (3.2), we have

$$0<\frac{f(x^p)}{p(p-1)G(q^p)^{p-4}}\leq-\frac{\theta}{3}+O(p^{-1}),$$

which is a contradiction for large $p>n$.

Next, suppose that $0\notin\hat{I}$ and $\hat{I}\cap I_2\neq\emptyset$. Then, using (3.6), (3.7), and (3.8) in (3.2) again, we get the same contradiction as above.

Finally, if we suppose that $0\notin\hat{I}$ and $\hat{I}\cap I_2=\emptyset$, then $\emptyset\neq\hat{I}\subset I_3$.

Hence, using (3.6), (3.7), and (3.8) in (3.2), we have

$$0\leq F^+_{\hat{I}}(q^p,X^p)+O(p^{-1}).$$

Thus, sending $p\to\infty$ in the above, by Lemma 3.2 we get

$$0\leq F^+_{\hat{I}}(q,X)\leq(F^+_\infty)^*(q,X),$$

which is a contradiction to (3.3).

Next, we shall verify that u is a subsolution of

$$\min\{G(Du)-1,F^-_\infty(Du,D^2u)\}\leq 0\quad\text{in }\Omega.$$

Suppose that, for $\phi\in C^2(\Omega)$, $u(\hat{x})=\phi(\hat{x})$ for $\hat{x}\in\Omega$, $u(x)<\phi(x)$ for $x\in\Omega\setminus\{\hat{x}\}$, and

$$\lim_{\varepsilon\to 0}\inf\left\{\min\{G(q')-1,F^-_\infty(q',X')\}\mid|q'-q|+|X'-X|<\varepsilon\right\}\geq\theta\tag{3.9}$$

for some $\theta>0$, where $q=D\phi(\hat{x})$ and $X=D^2\phi(\hat{x})$. We will get a contradiction again.

As before, we may choose $x^p\in\Omega$ such that $u^p-\phi$ attains its maximum at x^p and $\lim_{p\to\infty}x^p=\hat{x}$. By Proposition 2.3, we see that

$$\begin{aligned}\frac{f(x^p)}{p(p-2)}\ \geq\ &-|q^p_0|^{p-4}\sum_{k,l\in I_1}\left(\frac{\delta_{kl}}{p-2}(q^p_k)^2+q^p_kq^p_l\right)X^p_{kl}\\ &-\frac{p-1}{p-2}\left(\sum_{k\in I_2}(q^p_k)^{p-2}X^p_{kk}-\min_{k\in I_3}(-(q^p_k)^{p-2}X^p_{kk})\right)\\ =:&\ A_1+A_2+A_3,\end{aligned}\tag{3.10}$$

where $(q^p, X^p) = (D\phi(x^p), D^2\phi(x^p))$.

In view of (3.10), we see that $G(q^p) > 1$ for large $p > n$ since $G(q) > 1$. Moreover, we have

$$(F_\infty^-)_*(q, X) \geq \theta. \tag{3.11}$$

Following the argument for supersolutions with (2) of Proposition 3.3, we can estimate $\hat{A}_k := A_k/G(q^p)^{p-4}$ from below in the following manner:

$$\hat{A}_1 \geq \begin{cases} \dfrac{\theta}{3}, & \text{provided } 0 \in \hat{I}, \\ O(p^{-1}), & \text{provided } 0 \notin \hat{I}, \end{cases}$$

$$\hat{A}_2 \geq \begin{cases} \dfrac{\theta}{3}, & \text{provided } \hat{I} \cap I_2 \neq \emptyset, \\ O(p^{-1}), & \text{provided } \hat{I} \cap I_2 = \emptyset, \end{cases}$$

and

$$\hat{A}_3 \geq \begin{cases} \dfrac{\theta}{3}, & \text{provided } \hat{I} \cap I_3 = I_3, \\ O(p^{-1}), & \text{provided } \hat{I} \cap I_3 \neq I_3. \end{cases}$$

Here we use the same nonempty set $\hat{I} \subset I[q]$ as before.

If $0 \in \hat{I}$ or if $0 \notin \hat{I}$ and $\hat{I} \cap I_2 \neq \emptyset$, then, using these estimates in (3.10), we have

$$\frac{f(x^p)}{p(p-2)G(q^p)^{p-4}} \geq \frac{\theta}{3} + O(p^{-1}),$$

which is a contradiction for large $p > n$ since $G(q^p) > 1$.

On the other hand, in the case when $0 \notin \hat{I}$ and $\hat{I} \cap I_2 = \emptyset$, we see that $\emptyset \neq \hat{I} \subset I_3$. Thus, since $0 \geq F_{\hat{I}}^-(q^p, X^p) + O(p^{-1})$, Lemma 3.2 yields that

$$0 \geq F_{\hat{I}}^-(q, X) \geq (F_\infty^-)_*(q, X),$$

which contradicts (3.11). □

3.2. Proof of Lemma 3.2 and Proposition 3.3.

Proof of Lemma 3.2. We give only a proof of the representation formula for $(F_\infty^+)^*$ since we can show the other assertion similarly.

We may suppose that $G(q) > 0$ since the assertion is trivial if $G(q) = 0$; $q = 0$ by (A1).

For $(q, X) \in \mathbf{R}^n \times S^n$, we fix any nonempty set $J \subset I[q]$. We shall see that

$$(F_\infty^+)^*(q, X) \geq -\sum_{k,l \in I_1} 1_J q_k q_l X_{kl} - \sum_{k \in I_2 \cap J} q_k^2 X_{kk} + F_{I_3 \cap J}^+(q, X).$$

Case 1: $0 \in J$. First, we shall consider the case when $I_1 \cap (I_2[q] \cup I_3[q]) \neq \emptyset$.

In this case, we note that there is $k \in I_1$ such that $\{k\} = I_1 \cap (I_2[q] \cup I_3[q])$ and $q_j = 0$ for $j \in I_1 \setminus \{k\}$. Thus, taking $q^\varepsilon = (q_1^\varepsilon, \ldots, q_n^\varepsilon)$, where

$$q_j^\varepsilon = \begin{cases} q_j, & \text{provided } j \in I_1 \cup (J \cap I), \\ q_j - \varepsilon \operatorname{sgn}(q_j) & \text{otherwise,} \end{cases}$$

we have

$$\lim_{\varepsilon\downarrow 0} F_\infty^+(q^\varepsilon, X) = -\sum_{k,l\in I_1} q_k q_l X_{kl} - \sum_{j\in I_2\cap J} q_j^2 X_{jj} + F^+_{I_3\cap J}(q,X). \tag{3.12}$$

Even when $I_1 \cap (I_2[q] \cup I_3[q]) = \emptyset$, using the same q^ε as above, we get (3.12).

Case 2: $0 \notin J$. Let $q^\varepsilon = (q_1^\varepsilon, \dots, q_n^\varepsilon)$ be given by

$$q_k^\varepsilon = \begin{cases} q_k - \varepsilon \operatorname{sgn}(q_k), & \text{provided } k \in I \setminus J, \\ q_k, & \text{provided } k \in J. \end{cases}$$

Noting that $I[q^\varepsilon] = J$ for small $\varepsilon > 0$, we have

$$\lim_{\varepsilon\downarrow 0} F_\infty^+(q^\varepsilon, X) = -\sum_{k\in I_2\cap J} q_k^2 X_{kk} + F^+_{I_3\cap J}(q,X).$$

In order to show the opposite inequality, we choose $(q^m, X^m) \in \mathbf{R}^n \times S^n$ so that

$$\lim_{m\to\infty} (q^m, X^m) = (q, X) \quad \text{and} \quad \lim_{m\to\infty} F_\infty^+(q^m, X^m) = (F_\infty^+)^*(q,X).$$

By taking a subsequence if necessary, we may suppose that there is a nonempty set $J \subset \{0\} \cup I_2 \cup I_3$ such that $J = I[q^p]$ for all $p > n$. Hence we have

$$\begin{aligned}(F_\infty^+)^*(q,X) &= \lim_{m\to\infty} F_\infty^+(q^m, X^m) \\ &= -\sum_{k,l\in I_1} 1_J q_k q_l X_{kl} - \sum_{k\in I_2\cap J} q_k^2 X_{kk} + F^+_{I_3\cap J}(q,X). \qquad \square\end{aligned}$$

Next, using Lemma 3.2, we shall prove Proposition 3.3.

Proof of Proposition 3.3. We give only a proof for assertions of (1) since those for (2) can be proved similarly.

We introduce the notation: for $k \in I$, $\varepsilon > 0$, and $q = (q_1, \dots, q_n) \in \mathbf{R}^n$, we set

$$q_j^{k,\varepsilon} = \begin{cases} q_k & \text{for } j = k, \\ q_j - \varepsilon \operatorname{sgn}(q_j) & \text{otherwise.} \end{cases}$$

Proof of (i). Assume that $0 \in I[q]$. As noted in the proof of Lemma 3.2, we see that if $I_1 \cap I_j[q] \neq \emptyset$ for $j = 2$ or 3, then $q_j = 0$ for $j \in I_1 \setminus \{k\}$. Moreover, $I_1 \cap I_j[q]$ consists of a single point. Thus we remark that if $I_1 \cap I_j[q] = \{k\}$ for $j = 2$ or 3, then

$$-\sum_{j,l\in I_1} q_j q_l X_{jl} = -q_k^2 X_{kk}.$$

Thus we easily see that

$$-\theta \geq \lim_{\varepsilon\downarrow 0} F_\infty^+(q^{k,\varepsilon}, X) \geq \begin{cases} -2\displaystyle\sum_{k,l\in I_1} q_k q_l X_{kl}, & \text{provided } I_1 \cap I_2[q] \neq \emptyset, \\ -\displaystyle\sum_{k,l\in I_1} q_k q_l X_{kl}, & \text{provided } I_1 \cap I_2[q] = \emptyset \\ & \text{and } I_1 \cap I_3[q] \neq \emptyset. \end{cases}$$

In the remaining case (i.e., $I_1 \cap I_j[q] = \emptyset$ for $j = 2, 3$), we have

$$-\theta \geq \lim_{\varepsilon\downarrow 0} F_\infty^+(q^{0,\varepsilon}, X) = -\sum_{k,l\in I_1} q_k q_l X_{kl},$$

where $q^{0,\varepsilon} = (q_1^{0,\varepsilon}, \dots, q_n^{0,\varepsilon})$ is given by

$$q_k^{0,\varepsilon} = \begin{cases} q_k, & \text{provided } k \in I_1, \\ q_k - \varepsilon \mathrm{sgn}(q_k), & \text{provided } k \in I \setminus I_1. \end{cases}$$

Proof of (ii). If there is $k \in I_3[q] \setminus I_2[q]$, then as before, we have

$$-\theta \geq \lim_{\varepsilon \to 0} F_\infty^+(q^{k,\varepsilon}, X) = (-q_k^2 X_{kk}) \vee 0 \geq 0.$$

Thus we may suppose that $I_3[q] \setminus I_2[q] = \emptyset$.

Choose $k \in I_2[q]$. Then, as before, we have

$$-\theta \geq \lim_{\varepsilon \to 0} F_\infty^+(q^{k,\varepsilon}, X) \geq -q_k^2 X_{kk}. \qquad \square$$

4. Comparison principle. When we try to establish the comparison principle, we immediately meet some difficulties.

First, $F_\infty^\pm$ contains serious discontinuity with respect to Du-variables. Moreover, if $I_3 \neq \emptyset$, then F_∞^+ does not coincide with F_∞^-. Thus, to our knowledge, we cannot apply the standard argument in the theory of viscosity solutions to show that the comparison principle holds for sub- and supersolutions of the limit PDEs.

Therefore, we will impose the local convexity of subsolutions or the local concavity of supersolutions for our comparison result below. In the next section, we will present a sufficient condition for the existence of (locally) concave solutions.

As explained in Example 1.2, if $I_3 = I$ and $I_1 = I_2 = \emptyset$, then we cannot expect that the comparison principle holds in general.

To avoid this difficulty, we suppose that

$$\text{(A2)} \qquad I_3 \subset I_2.$$

We note that (A1) is not necessary for the comparison result below, but we have to suppose that

$$\text{(A3)} \qquad I_1 \cup I_2 \neq \emptyset.$$

THEOREM 4.1. *Assume that* (A2) *and* (A3) *hold. Let u and $v \in C(\overline{\Omega})$ be, respectively, a subsolution and a supersolution of*

$$\min\{G(Du) - 1, F_\infty^-(Du, D^2u)\} \leq 0 \quad \textit{in } \Omega$$

and

$$\min\{G(Dv) - 1, F_\infty^+(Dv, D^2v)\} \geq 0 \quad \textit{in } \Omega.$$

If we assume that either u or $-v$ is locally convex, then we have

$$\max_{\overline{\Omega}}(u - v) \leq \max_{\partial\Omega}(u - v).$$

Remark. In our proof, we modify Jensen's argument in [9] for the case when $I_2 \cup I_3 = \emptyset$ since the limit PDEs are possibly discontinuous.

We give two approximations for sub- and supersolutions in Lemma 4.2. The approximations in Lemma 4.2 were obtained in [9], but our lemma contains more precise information.

To demonstrate Theorem 4.1, we will use the approximation of supersolutions and a property from Lemma 4.2. However, we can prove our comparison result more easily using both approximations with no use of the precise information. See the remark after the proof.

Lemma 4.2. *Under the assumptions in Theorem* 4.1, *for any* $\varepsilon > 0$, *there are functions* $\bar{u}$ *and* $\bar{v} \in C(\overline{\Omega})$ *and a constant* $\tau > 0$ *satisfying the following properties:*

(i) $\max_{\overline{\Omega}}(|u-\bar{u}| + |v-\bar{v}|) < \varepsilon$.

(ii) $\bar{u}$ *and* $\bar{v}$ *are, respectively, a subsolution and a supersolution of*

$$\min\{G(D\bar{u}) - 1, F_\infty^-(D\bar{u}, D^2\bar{u})\} + \tau \le 0 \tag{4.1}$$

and

$$\min\{G(D\bar{v}) - 1, F_\infty^+(D\bar{v}, D^2\bar{v})\} - \tau \ge 0. \tag{4.2}$$

(iii) *If* u (*resp.,* v) *is convex* (*resp., concave*), *then so is* $\bar{u}$ (*resp.,* $\bar{v}$).

(iv) *For any* $x \in \Omega$, *there are constants* $\alpha_k, \beta_k \ge \tau$ ($k = 1, 2$) *such that for any* $(q, X) \in \overline{J}^{2,+}\bar{u}(x)$ (*resp.,* $(q, X) \in \overline{J}^{2,-}\bar{v}(x)$), *there is* $(\bar{q}, \bar{X}) \in \overline{J}^{2,+}u(x)$ (*resp.,* $(\bar{q}, \bar{X}) \in \overline{J}^{2,-}v(x)$) *satisfying the following property:*

$$\bar{q} = \alpha_1 q \text{ and } \bar{X} = \alpha_1 X - \beta_1 q \otimes q \quad (\text{resp., } \bar{q} = \alpha_2 q \text{ and } \bar{X} = \alpha_2 X + \beta_2 q \otimes q).$$

Proof of Theorem 4.1. Suppose that there is $\theta > 0$ such that

$$\max_{\overline{\Omega}}(u - v) - \max_{\partial\Omega}(u - v) \ge 2\theta.$$

In view of Lemma 4.2, we choose a supersolution $\bar{v} \in C(\overline{\Omega})$ (for small $\varepsilon > 0$) of (4.2) with a positive constant $\tau > 0$. Thus we may suppose that

$$\max_{\overline{\Omega}}(u - \bar{v}) - \max_{\partial\Omega}(u - \bar{v}) \ge \theta. \tag{4.3}$$

We set $\Phi(x, y) = u(x) - \bar{v}(y) - \alpha|x - y|^2/2$. Because of (4.3), by the standard argument (see [4], for instance), we may find $(x^\alpha, y^\alpha) \in \Omega \times \Omega$ such that

$$\max_{\overline{\Omega}\times\overline{\Omega}} \Phi = \Phi(x^\alpha, y^\alpha) \quad \text{and} \quad \lim_{\alpha\to\infty}(x^\alpha, y^\alpha) = (z, z),$$

where $z \in \Omega$ satisfies $\max_{\overline{\Omega}}(u - \bar{v}) = (u - \bar{v})(z)$.

In view of Theorem 3.2 in [4], for instance, we find X^α and $Y^\alpha \in S^n$ such that

$$\begin{pmatrix} X^\alpha & 0 \\ 0 & -Y^\alpha \end{pmatrix} \le 3\alpha \begin{pmatrix} I & -I \\ -I & I \end{pmatrix}, \tag{4.4}$$

$$(\alpha(x^\alpha - y^\alpha), X^\alpha) \in \overline{J}^{2,+}u(x^\alpha) \quad \text{and} \quad (\alpha(x^\alpha - y^\alpha), Y^\alpha) \in \overline{J}^{2,-}\bar{v}(y^\alpha).$$

We note that (4.4) implies

$$X^\alpha \le Y^\alpha. \tag{4.5}$$

Hence, setting $q^\alpha = \alpha(x^\alpha - y^\alpha)$, we have

$$\lim_{\varepsilon\to 0} \sup\left\{\min\{G(q) - 1, F_\infty^+(q, Y)\} \;\middle|\; |q - q^\alpha| + |Y - Y^\alpha| < \varepsilon\right\} \ge \tau \tag{4.6}$$

and

$$\lim_{\varepsilon\to 0}\inf\left\{\min\{G(q)-1, F_\infty^-(q,X)\}\ \big|\ |q-q^\alpha|+|X-X^\alpha|<\varepsilon\right\}\le 0. \tag{4.7}$$

In view of (4.6), we find a sequence $(q^m, Y^m)\in \mathbf{R}^n\times S^n$ such that $\lim_{m\to\infty}(q^m, Y^m)=(q^\alpha, Y^\alpha)$ and that

$$\lim_{m\to\infty}\min\{G(q^m)-1, F_\infty^+(q^m,Y^m)\}\ge\tau. \tag{4.8}$$

Since G is continuous, we have

$$G(q^\alpha)-1\ge\tau. \tag{4.9}$$

Let us first consider the case when u is locally convex.

Lemma 2.2, together with (4.5), yields that $Y^\alpha\ge 0$ since u is locally convex. Hence, if $G(q^m)=q_0^m$ (≥ 1) for $m\ge 1$, then we have

$$\lim_{m\to\infty}F_\infty^+(q^m,Y^m)\le-\lim_{m\to\infty}\sum_{k,l\in I_1}q_k^m q_l^m Y_{kl}^m\le 0,$$

which contradicts (4.8).

Thus we may suppose that $G(q^m)>q_0^m$ for all $m\ge 1$. In this case, due to (A3), we find $k\in I_2$ such that $G(q^m)=|q_k^m|\ge 1$. Hence we have

$$\lim_{m\to\infty}F_\infty^+(q^m,Y^m)\le 0,$$

which contradicts (4.8) again.

Next, let us suppose that v is locally concave.

In view of (iv) in Lemma 4.2, we find $(\bar q,\bar Y)\in\overline{J}^{2,-}v(y^\alpha)$ such that $\bar q=\alpha_2 q^\alpha$ and

$$\bar Y=\alpha_2 Y^\alpha+\beta_2 q^\alpha\otimes q^\alpha\quad\text{for some }\alpha_2,\beta_2\ge\tau.$$

Hence, since $G(q^\alpha)\ge 1$, by Lemma 2.2, we find $\sigma>0$ such that

$$\langle X^\alpha q^\alpha, q^\alpha\rangle\le\langle Y^\alpha q^\alpha, q^\alpha\rangle\le-\sigma. \tag{4.10}$$

Now we choose $(\hat q^m, X^m)\in\mathbf{R}^n\times S^n$ such that $\lim_{m\to\infty}(\hat q^m, X^m)=(q^\alpha, X^\alpha)$ and that

$$\lim_{m\to\infty}\min\{G(\hat q^m)-1, F_\infty^-(\hat q^m, X^m)\}\le 0. \tag{4.11}$$

We note that (4.10) implies

$$\langle X^m\hat q^m,\hat q^m\rangle\le-\frac{\sigma}{2}. \tag{4.12}$$

Since we may suppose $G(\hat q^m)-1>0$ by (4.9), (4.11) yields that

$$\lim_{m\to\infty}F_\infty^-(\hat q^m, X^m)\le 0.$$

However, as before, by (A2), this implies a contradiction to (4.12). ∎

Remark. If we also approximate u by $\bar u$ by Lemma 4.2, then (4.7) becomes

$$\lim_{\varepsilon\to 0}\inf\{\min\{G(q)-1, F_\infty^-(q,X)\}\ |\ |q-q^\alpha|+|X-X^\alpha|<\varepsilon\}\le-\tau. \tag{4.7$'$}$$

Then, in the case when v is locally concave, we directly get a contradiction from (4.7′) with no use of (iv) of Lemma 4.2.

Since in the proof of Theorem 4.1 we need only to suppose that v is a supersolution of

$$G(Du) - 1 \geq 0 \quad \text{in } \Omega,$$

we have the following corollary, which will be useful to show that u^p converges to the corresponding distance function from $\partial\Omega$ in a typical example.

COROLLARY 4.3. *Assume that* (A2) *and* (A3) *hold. Let u and $v \in C(\overline{\Omega})$ be, respectively, a subsolution and a supersolution of*

$$\min\{G(Du) - 1, F_\infty^-(Du, D^2u)\} \leq 0 \quad \textit{in } \Omega$$

and

$$G(Dv) - 1 \geq 0 \quad \textit{in } \Omega.$$

If we assume that v is locally concave, then we have

$$\max_{\overline{\Omega}}(u - v) \leq \max_{\partial\Omega}(u - v).$$

Remark. If u is a subsolution of $G(Du) - 1 \leq 0$ in the above, then we do not need to suppose that u (or $-v$) is locally convex to show the assertion. See the proof of the comparison principle for the eikonal equation in [8].

Proof of Lemma 4.2. Set $C_0 := \max_{\overline{\Omega}}(|u| + |v|) + 1$. We shall first construct $\bar{u}$. We set $w = u + \delta u^2$, where $\delta := \min\{1/(4C_0), \varepsilon/C_0^2\}$. We note that $-3/(16\delta) \leq w$. We notice that w is convex if u is convex. Thus, because we will take $\bar{u} = \rho w$ for some $\rho > 0$, we verify that (iii) holds.

Suppose that for $\phi \in C^2(\Omega)$, $w(\hat{x}) = \phi(\hat{x})$ for $\hat{x} \in \Omega$, and $w \leq \phi$ in Ω. We may suppose that $-3/(16\delta) \leq \phi$.

Choose $\psi \in C^2(\Omega)$ such that $|\psi| \leq C_0$, $\phi = \psi + \delta\psi^2$, and $u - \psi$ attains its maximum at x^0. Hence we can find $(q^m, X^m) \in \mathbf{R}^n \times S^n$ such that

$$\lim_{m\to\infty}(q^m, X^m) = (D\psi(\hat{x}), D^2\psi(\hat{x}))$$

and

$$\min\{G(q^m) - 1, F_\infty^-(q^m, X^m)\} \leq \frac{1}{m}.$$

We note that

$$D\psi = \frac{D\phi}{1 + 2\delta\psi} \quad \text{and} \quad D^2\psi = \frac{D^2\phi}{1 + 2\delta\psi} - \frac{2\delta D\phi \otimes D\phi}{(1 + 2\delta\psi)^3}.$$

If there is a subsequence of $\{q^m\}$ (denoted by the same symbol) such that

$$G(q^m) - 1 \leq \frac{1}{m},$$

then we have

$$G(D\phi(\hat{x})) \leq \frac{1}{1 + 2\delta\psi(\hat{x})} \leq \frac{1}{1 - \delta C_0}. \tag{4.13}$$

On the other hand, if there is no such a subsequence of $\{q^m\}$, then we may suppose that

$$G(q^m) \geq 1,$$

and, moreover,

$$-\sum_{k,l\in I_1[q^m]} q_k^m q_l^m X_{kl}^m - \sum_{k\in I_2[q^m]} (q_k^m)^2 X_{kk}^m + f_{I_3[q^m]}^-(q^m, X^m) \leq \frac{1}{m}.$$

Setting $\alpha = 1 + 2\delta\psi(\hat{x}) \in [1/2, 2]$, $\hat{q}^m = \alpha q^m$, and $\hat{X}^m = \alpha X^m + 2\delta\alpha^{-2}\hat{q}^m \otimes \hat{q}^m$, we have

$$\begin{aligned}\frac{8}{m} \geq & -\sum_{k,l\in I_1[\hat{q}^m]} \hat{q}_k^m \hat{q}_l^m \left(\hat{X}_{kl}^m - 2\delta\alpha^{-2}\hat{q}_k^m \hat{q}_l^m\right) \\ & - \sum_{k\in I_2[\hat{q}^m]} (\hat{q}_k^m)^2 \left(\hat{X}_{kk}^m - 2\delta\alpha^{-2}(\hat{q}_k^m)^2\right) + f_{I_3[q^m]}^-(q^m, X^m).\end{aligned}$$

Hence, in the case when $G(\hat{q}^m) = \hat{q}_0^m$, we can choose a constant $\hat{\tau} > 0$ such that for any $m \geq 1$,

$$\begin{aligned}-\hat{\tau} + \frac{8}{m} \geq & -\sum_{k,l\in I_1} \hat{q}_k^m \hat{q}_l^m \hat{X}_{kl}^m - \sum_{k\in I_2[\hat{q}^m]} (\hat{q}_k^m)^2 \hat{X}_{kk}^m + f_{I_3[\hat{q}^m]}^-(\hat{q}^m, \hat{X}^m) \\ = & F_\infty^-(\hat{q}^m, \hat{X}^m).\end{aligned}$$

If $G(\hat{q}^m) > \hat{q}_0^m$, then, noting (A2), we find $k \in I_2$ such that $G(\hat{q}^m) = |\hat{q}_k^m|$. Thus we can find a constant $\hat{\tau} > 0$ satisfying the above inequality. (Notice that the first term does not exist in this case.)

Finally, in view of (4.13), taking $\bar{u} = \rho w$, where $\rho = 1 - 2\delta C_0$, we verify easily that $\bar{u}$ is a subsolution of

$$\min\left\{G(D\bar{u}) - 1 + \frac{\delta C_0}{1 - 2\delta C_0}, F_\infty^-(D\bar{u}, D^2\bar{u}) + \hat{\tau}(1 - 2\delta C_0)^3\right\} \leq 0 \quad \text{in } \Omega.$$

Therefore, taking $\tau := \min\{\delta C_0/(1 - 2\delta C_0), \hat{\tau}(1 - 2\delta C_0)^3\}$, we verify that (4.1) holds for $\bar{u}$.

In order to check (iv), we let $(q, X) \in \overline{J}^{2,+}\bar{u}(x)$ for $x \in \Omega$. Choose $(q^m, X^m) \in J^{2,+}\bar{u}(x^m)$ so that $\lim_{m\to\infty}(x^m, u(x^m), q^m, X^m) = (x, u(x), q, X)$. For each $m \geq 1$, we can choose $\phi^m \in C^2(\Omega)$ such that $\bar{u} \leq \phi^m$ in Ω, $\bar{u}(x^m) = \phi^m(x^m)$, and $(D\phi^m(x^m), D^2\phi(x^m)) = (q^m, X^m)$.

We can select $\psi^m \in C^2(\Omega)$ such that $\phi^m = \rho\psi^m(1 + \delta\psi^m)$, $u \leq \psi^m$ in Ω, $u(x^m) = \psi^m(x^m)$, and $(D\psi^m(x^m), D^2\psi(x^m)) \in J^{2,+}u(x^m)$. Hence we have

$$\hat{q}^m := D\psi^m(x^m) = \frac{q^m}{\rho(1 + 2\delta\psi(x^m))}$$

and

$$X^m := D^2\psi^m(x^m) = \frac{X^m}{\rho(1 + 2\delta\psi^m(x^m))} - \frac{2\delta q^m \otimes q^m}{\rho(1 + 2\delta\psi^m(x^m))^3}.$$

By taking $(\bar{q}, \bar{X}) = \lim_{m\to\infty}(\hat{q}^m, \hat{X}^m)$ with $\alpha_1 = \rho^{-1}(1 + 2\delta u(x))^{-1}$ and $\beta_1 = 2\delta\rho^{-1}(1 + 2\delta u(x))^{-3}$, we get (iv) for smaller $\tau > 0$ if necessary.

To construct $\bar{v}$, at the first stage in the above, we use the transformation $\hat{w} = v - \delta v^2$. We can follow the argument for $\bar{u}$. We also refer to [9]. □

5. Existence of concave solutions. In the case when $I_3 \neq \emptyset$, we do not know nice approximate PDEs, which implies the "power concavity" of solutions. Therefore, in this section, we suppose that

$$\text{(A4)} \qquad I_3 = \emptyset.$$

We also suppose that

$$\text{(A5)} \qquad \Omega \quad \text{is convex.}$$

THEOREM 5.1 (cf. Theorem 2 in [16]). *Assume that* (A1), (A4), *and* (A5) *hold. Let* $u \in W_0^{1,p}(\Omega)$ *be the* (*unique*) *minimizer of*

$$\inf\left\{ \|Dv\|^p - \int_\Omega v dx \;\Big|\; v \in W_0^{1,p}(\Omega) \right\}.$$

Then $u^{\frac{p-1}{p}}$ *is concave for* $p > n$.

Remark. For the conclusion in Theorem 5.1, we do not have to suppose $p > n$. However, for the existence of concave solutions of the limit PDEs, we need only the power concavity for large $p > 1$.

Since $(u^p)^{(p-1)/p}$ converges $\lim_{p\to\infty} u^p$ uniformly in $\overline{\Omega}$, where u^p is the unique minimizer of the above variational problem, Theorem 5.1 yields the following.

COROLLARY 5.2. *Assume that* (A1), (A4), *and* (A5) *hold. Then there exists a unique concave solution* $u \in W^{1,\infty}(\Omega)$, *which satisfies that* $u = 0$ *on* $\partial\Omega$ *of*

$$\begin{cases} \min\{G(Du) - 1, F_\infty^-(Du, D^2u)\} \leq 0 & \textit{in } \Omega, \\ \min\{G(Du) - 1, F_\infty^+(Du, D^2u)\} \geq 0 & \textit{in } \Omega. \end{cases}$$

Since our proof is a modification of that by Sakaguchi in [16], we give only a sketch of proof.

Sketch of proof of Theorem 5.1. For $\varepsilon > 0$, we shall consider the following regularized variational problem:

$$\text{(5.1)} \qquad \inf\left\{ \int_\Omega g_\varepsilon^p(v, Dv)dx - \int_\Omega v dx \;\Big|\; v \in W_0^{1,p}(\Omega) \right\},$$

where

$$g_\varepsilon^p(r,p) = \left(\varepsilon|r|^{\frac{2}{p}} + |P_1[p]|^2\right)^{\frac{p}{2}} + \sum_{k\in I_2} \left(\varepsilon|r|^{\frac{2}{p}} + |p_k|^2\right)^{\frac{p}{2}}.$$

It is easy to see that a (unique) minimizer $u_\varepsilon \in W_0^{1,p}(\Omega)$ of (5.1) exists. Moreover, by the standard argument (see [16] for example), we see that

$$\text{(5.2)} \qquad \begin{cases} \text{(i)} & u_\varepsilon \to u \text{ in } W^{1,p}(\Omega) \text{ as } \varepsilon \to 0, \\ \text{(ii)} & u_\varepsilon \geq 0 \text{ in } \Omega, \end{cases}$$

where $u \in W_0^{1,p}(\Omega)$ is the unique minimizer of the variational problem (5.1) with $\varepsilon = 0$.

According to the weak Harnack inequality in [17], we see that

$$\text{(5.3)} \qquad u > 0 \quad \text{in } \Omega.$$

Hence (5.3) and the uniform convergence of u_ε to u imply that for any $\delta > 0$, there are $\varepsilon(\delta), \tau(\delta) > 0$ such that

$$u_\varepsilon \geq \tau(\delta) \quad \text{in } \Omega_\delta \quad \text{for } \varepsilon \in (0, \varepsilon(\delta)), \tag{5.4}$$

where

$$\Omega_\delta = \{x \in \Omega \mid \operatorname{dist}(x, \partial\Omega) > \delta\}.$$

We may suppose that Ω_δ is strictly convex and $\partial\Omega_\delta$ is smooth.

From now on, for simplicity, we shall write u for u_ε. Also, we write $D_1 u$ for $P_1[Du]$.

It is obvious that u is a weak solution (in the distribution sense) of

$$-\sum_{k \in I_1} \left[\left(\varepsilon u^{\frac{2}{p}} + |D_1 u|^2 \right)^{\frac{p-2}{p}} u_{x_k} \right]_{x_k} - \sum_{k \in I_2} \left[\left(\varepsilon u^{\frac{2}{p}} + |u_{x_k}|^2 \right)^{\frac{p-2}{2}} u_{x_k} \right]_{x_k} = G_\varepsilon(u, Du),$$

where

$$G_\varepsilon(u, Du) = \frac{1}{p} - \frac{\varepsilon}{p} u^{\frac{2-p}{p}} \left\{ \left(\varepsilon u^{\frac{2}{p}} + |D_1 u|^2 \right)^{\frac{p-2}{p}} + \sum_{k \in I_2} \left(\varepsilon u^{\frac{2}{p}} + |u_{x_k}|^2 \right)^{\frac{p-2}{p}} \right\}.$$

We remark that $u \in C^\infty(\Omega_\delta)$ (see [16] for example). Thus, setting $v = u^{\frac{p-1}{p}}$, we observe that v is a classical solution of

$$-\sum_{k \in I_1} A(|D_1 v|)_{x_k} - \sum_{k \in I_2} A(v_{x_k})_{x_k} = \hat{G}_\varepsilon(v, Dv), \tag{5.5}$$

where for $r \in \mathbf{R}$,

$$A(r) = \left(\varepsilon + \left(\frac{p}{p-1} \right)^2 r^2 \right)^{\frac{p-2}{2}},$$

and for $(r, q) \in \mathbf{R} \times \mathbf{R}^n$,

$$\hat{G}_\varepsilon(r, q) = \frac{p-1}{vp^2} \left\{ 1 + \left(\frac{|P_1[q]|^2}{p-1} - \varepsilon \right) A(|P_1[q]|) + \sum_{k \in I_2} \left(\frac{q_k^2}{p-1} - \varepsilon \right) A(q_k) \right\}.$$

We note that the left-hand side of (5.5) can be written in the following way:

$$-\sum_{k,l \in I} a_\varepsilon^{kl}(Dv) v_{x_k x_l},$$

where $(a_\varepsilon^{kl}(q))$ is positive semidefinite.

In order to apply the concave maximum principle by Kennington in [13] to (5.5), we first remark that the right-hand side of (5.5) is positive in Ω_δ for small $\varepsilon > 0$.

Setting $w(x, y; t) = (1-t)v(x) + (1-t)v(y) - v((1-t)x + ty)$ for $x, y \in \Omega$ and $t \in (0, 1)$, we also remark that Korevaar's Lemma 2.1 in [14] yields that w does not attain its positive maximum on $\partial(\Omega_\delta \times \Omega_\delta) \times [0, 1]$. Therefore, Theorem 3.1 in [13] implies that $w \leq 0$ in $\overline{\Omega}_\delta \times \overline{\Omega}_\delta \times [0, 1]$.

Since $(u_\varepsilon)^{(p-1)/p}$ is concave in Ω_δ for small $\varepsilon > 0$, so is $u^{(p-1)/p}$ in Ω_δ. Therefore, $u^{(p-1)/p}$ is concave in the whole domain Ω. □

6. Some remarks and examples.

6.1. Convergence to a distance function. In this subsection, for simplicity, we shall consider only the case when

$$\text{(A6)} \qquad I_2 = I \quad \text{and} \quad I_1 = I_3 = \emptyset.$$

In contrast with the result in [12] and [3], we show that the minimizer u^p of the variational problem (2.2) converges to the following distance function from $\partial\Omega$:

$$d_1(x, \partial\Omega) = \inf \left\{ \sum_{k\in I} |x_k - y_k| \ \middle|\ y \in \partial\Omega \right\}.$$

We first give a remark which was informed by Ishii.

Proposition 6.1. *We see that* $d_1(\cdot, \partial\Omega)$ *is a supersolution of*

$$\text{(6.1)} \qquad G(Du) - 1 = 0 \quad \textit{in } \Omega.$$

Proof. Due to the well-known fact in the theory of viscosity solutions, it is enough to show that for each $y = (y_1, \ldots, y_n) \in \partial\Omega$,

$$h(x) := \sum_{k\in I} |x_k - y_k|$$

is a supersolution of (6.1).

For any $x^0 \in \Omega$, we let $p \in \mathbf{R}^n$ satisfy that $(p, X) \in J^{2,-}u(x^0)$ for some $X \in S^n$. We may suppose that $x^0 = 0$. We notice that $y \neq 0$. Thus we have

$$u(x) \geq u(0) + \langle p, x\rangle + o(|x|).$$

Hence it is easy to see that

$$|p_k| \leq 1 \quad \text{provided } y_k = 0 \quad \text{and} \quad |p_k| = 1 \quad \text{provided } y_k \neq 0.$$

Therefore, we verify that $G(p) = 1$. □

Next we prepare the following lemma.

Lemma 6.2. *Assume that* (A5) *holds. Then* $d_1(\cdot, \partial\Omega)$ *is concave in* Ω.

Proof. Fix $x^j = (x^j_1, \ldots, x^j_n) \in \Omega$ for $j = 1, 2$. Since $Q_j := \{y \in \mathbf{R}^n \mid \sum_{k\in I} |x^j_k - y_k| \leq d_1(x^j, \partial\Omega)\}$ for $j = 1, 2$ are in $\overline{\Omega}$, by (A5), we see that for $t \in [0, 1]$,

$$tQ_1 + (1-t)Q_2 = \{tx + (1-t)y \in \mathbf{R}^n \mid x \in Q_1,\ y \in Q_2\} \subset \overline{\Omega}.$$

Therefore, in view of a simple observation, we see that

$$d_1(tx^1 + (1-t)x^2, \partial\Omega) \geq td_1(x^1, \partial\Omega) + (1-t)d_1(x^2, \partial\Omega). \qquad □$$

We shall show that $d_1(\cdot, \partial\Omega)$ is indeed a solution of (6.1) when Ω is convex.

Proposition 6.3. *Assume that* (A5) *holds. Then* $d_1(\cdot, \partial\Omega)$ *is a solution of* (6.1).

Proof. In view of Proposition 6.1, it is sufficient to show that d_1 is a subsolution of (6.1).

For any $x^0 \in \Omega$, we fix $p \in \mathbf{R}^n$ such that $(p, X) \in J^{2,+}d_1(x^0, \partial\Omega)$ for some $X \in S^n$. We may suppose $x^0 = 0$.

We shall show that $G(p) \leq 1$. To this end, we consider the function $\phi(x) = -\sum_{k\in I} |x_k| + d_1(0, \partial\Omega)$ for $x \in \overline{\Omega}$.

We claim that ϕ is a subsolution of $G(D\phi) - 1 \leq 0$ in Ω. Indeed, for any $x \in \Omega$ and $\hat{p} \in \mathbf{R}^n$ such that $(\hat{p}, \hat{X}) \in J^{2,+}\phi(x)$ for some $\hat{X} \in S^n$, as in the proof of Proposition 6.2, we easily see that

$$|\hat{p}_k| \leq 1, \quad \text{provided } x_k = 0, \quad \text{and} \quad |\hat{p}_k| = 1, \quad \text{provided } x_k \neq 0.$$

Thus, noting that $\phi \leq 0$ on $\partial\Omega$, by Corollary 4.3 together with Proposition 6.1 and Lemma 6.2, we see that $\phi \leq d_1$ in $\overline{\Omega}$. Therefore, we have

$$\begin{aligned}\phi(x) \leq d_1(x, \partial\Omega) &\leq d_1(0, \partial\Omega) + \langle p, x\rangle + \frac{\langle Xx, x\rangle}{2} + o(|x|^2)\\ &= \phi(0) + \langle p, x\rangle + \frac{\langle Xx, x\rangle}{2} + o(|x|^2).\end{aligned}$$

Hence we have $(p, X) \in J^{2,+}\phi(0)$, which concludes the proof. □

Finally, combining Proposition 6.3 with Corollary 4.3, we obtain the following.

COROLLARY 6.4. *Assume that* (A5) *and* (A6) *hold. Assume also that*

$$\min_{x\in\overline{\Omega}} f(x) > 0.$$

Let u^p *be the minimizer of* (2.2).

Then we have

$$\lim_{p\to\infty} u^p(x) = d_1(x, \partial\Omega) \quad \textit{uniformly in } x \in \overline{\Omega}.$$

Proof. In view of Theorems 4.1 and 5.1, we may choose a subsequence u^{p_i} so that $u := \lim_{i\to\infty} u^{p_i}$ is a concave solution of

$$\min\left\{G(Du) - 1, -\sum_{k\in I[Du]} u_{x_k}^2 u_{x_k x_k}\right\} = 0 \quad \text{in } \Omega.$$

Hence Proposition 6.3 and Theorem 4.1 yield that $d_1 \leq u$ in $\overline{\Omega}$.

In view of Lemma 6.2 and Proposition 6.3, we see that d_1 is a concave supersolution of $G(Du) - 1 \geq 0$ in Ω. Thus Corollary 4.3 and Theorem 5.1 yield that $u \leq d_1$ in $\overline{\Omega}$. □

6.2. Examples. We consider three typical domains in $\mathbf{R}^2$; $I = \{1, 2\}$. For each domain, we deal with three cases:

$$\left\{\begin{aligned}\min\left\{|Du| - 1, -\sum_{k,l\in I} u_{x_k} u_{x_l} u_{x_k x_l}\right\} &= 0 \quad \text{in } \Omega,\\ u &= 0 \quad \text{on } \partial\Omega,\end{aligned}\right. \tag{6.2}$$

$$\left\{\begin{aligned}\min\left\{\max_{j\in I} |u_{x_j}| - 1, -\sum_{k\in I[Du(x)]} (u_{x_k})^2 u_{x_k x_k}\right\} &= 0 \quad \text{in } \Omega,\\ u &= 0 \quad \text{on } \partial\Omega,\end{aligned}\right. \tag{6.3}$$

and

$$(6.4)\qquad \begin{cases} \min\Big\{\max\limits_{j\in I}|u_{x_j}|-1, F^{+}_{I[Du]}(Du,D^2u)\Big\}\geq 0 & \text{in } \Omega,\\ \min\Big\{\max\limits_{j\in I}|u_{x_j}|-1, F^{+}_{I[Du]}(Du,D^2u)\Big\}\leq 0 & \text{in } \Omega,\\ \hfill u=0 & \text{on } \partial\Omega. \end{cases}$$

Here we use the notation $F^{\pm}_J$ for $J\subset I$ from section 3.

We notice that (6.2), (6.3), and (6.4), respectively, correspond to the cases when $I_2=I_3=\emptyset$, $I_1=I_3=\emptyset$, and $I_1=I_2=\emptyset$.

According to [12] and [3], we know that

$$\lim_{p\to\infty} u^p(x)=d(x,\partial\Omega)\quad\text{uniformly in } \overline{\Omega},$$

where u^p is the minimizer of the variational problem (2.2) for $I_2=I_3=\emptyset$.

Also, by Corollary 6.4, we see that

$$\lim_{p\to\infty} u^p(x)=d_1(x,\partial\Omega)\quad\text{uniformly in } \overline{\Omega},$$

where u^p is the minimizer of the variational problem (2.2) for $I_1=I_3=\emptyset$.

In Examples 6.1–6.3 below, although $d_1(\cdot,\partial\Omega)$ is a solution of (6.4), it does not imply that the corresponding u^p converges to $d_1(\cdot,\partial\Omega)$ because we do not know the uniqueness of (concave) solutions of (6.4).

Example 6.1. Let us consider $\Omega=\{x=(x_1,x_2)\in\mathbf{R}^2 \mid |x_k|<1 \text{ for } k\in I\}$.

Then we easily see that the following function is a solution of (6.2), (6.3), and (6.4):

$$u(x_1,x_2)=\begin{cases} 1-|x_1| & \text{for } |x_2|\leq |x_1|,\\ 1-|x_2| & \text{otherwise.}\end{cases}$$

Notice that $u(x)=d(x,\partial\Omega)=d_1(x,\partial\Omega)$.

We note that u is the unique solution of (6.2) and that it is the unique concave solution of (6.3).

Example 6.2. We next deal with $\Omega=\{(x_1,x_2)\in\mathbf{R}^2 \mid |x_1|+|x_2|<1\}$.

It is easy to see that the distance function $d(\cdot,\partial\Omega)$ is given by

$$d_1(x,\partial\Omega)=1-|x_1|-|x_2|\quad\text{for } x=(x_1,x_2)\in\Omega$$

and that it is the unique concave solution of (6.3).

Also, we see that the distance function $d(x,\partial\Omega)=d_1(x,\partial\Omega)/\sqrt{2}$ is the unique solution of (6.2).

Example 6.3. We shall treat $\Omega=B_1$. It is immediately seen that the unique solution of (6.2) is given by the distance function

$$d(x,\partial\Omega)=1-|x|\quad\text{for } x\in\Omega.$$

By an elementary calculation, we can show that the distance function $d_1(\cdot,\partial\Omega)$ is given by

$$d_1(x,\partial\Omega)=\begin{cases} \sqrt{1-x_2^2}-|x_1| & \text{for } |x_2|\leq|x_1|,\\ \sqrt{1-x_1^2}-|x_2| & \text{otherwise.}\end{cases}$$

Hence, by Corollary 6.4, we see that $d_1(\cdot, \partial\Omega)$ is the unique concave solution of (6.3).

We note that for $r \in (0,1)$, the level set $\Omega(r) := \{x \in \mathbf{R}^2 \mid d_1(x, \partial\Omega) > r\}$ strictly contains B_r. More precisely, we have

$$\Omega(r) = \bigcap_{k=1}^{4} B_1(\hat{x}^k),$$

where

$$\hat{x}^1 = (r-1, 0), \quad \hat{x}^2 = (0, r-1), \quad \hat{x}^3 = (1-r, 0), \quad \text{and} \quad \hat{x}^4 = (0, 1-r).$$

Acknowledgments. The authors thank Professor H. Ishii for his suggestions on the first draft.

REFERENCES

[1] G. Aronsson, *Extension of functions satisfying Lipschitz conditions*, Ark. Mat., 6 (1967), pp. 551–561.

[2] G. Aronsson, *On the partial differential equation $u_x^2 u_{xx} + 2u_x u_y u_{xy} + u_y^2 u_{yy} = 0$*, Ark. Mat., 7 (1968), pp. 395–425.

[3] T. Bhattacharya, E. DiBenedetto, and J. Manfredi, *Limits as $p \to \infty$ of $\triangle_p u_p = f$ and related extremal problems*, Rend. Sem. Mat. Univ. Politec. Torino, 1989 (1991), pp. 15–68.

[4] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[5] L. C. Evans, *Estimates for smooth absolutely minimizing Lipschitz extensions*, 1993 (1993), pp. 1–9.

[6] N. Fukagai, M. Ito, and K. Narukawa, *Limit as $p \to \infty$ of p-Laplace eigenvalue problems and L^∞-inequality of the Poincaré type*, Differential Integral Equations, 12 (1999), pp. 183–206.

[7] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, New York, 1983.

[8] H. Ishii, *A simple direct proof for uniqueness for solutions of the Hamilton-Jacobi equations of eikonal type*, Proc. Amer. Math. Soc., 100 (1987), pp. 247–251.

[9] R. Jensen, *Uniqueness of Lipschitz extensions: Minimizing the sup norm of the gradient*, Arch. Rational Mech. Anal., 123 (1993), pp. 51–74.

[10] P. Juutinen, *Minimization problems for Lipschitz functions via viscosity solutions*, Ann. Acad. Sci. Fenn. Math. Diss., 115 (1998).

[11] P. Juutinen, P. Lindqvist, and J. J. Manfredi, *The ∞-eigenvalue problem*, Arch. Rational Mech. Anal., 148 (1999), pp. 89–105.

[12] B. Kawohl, *On a family of torsional creep problems*, J. Reine Angew. Math., 410 (1990), pp. 1–22.

[13] A. U. Kennington, *Power concavity and boundary value problems*, Indiana Univ. Math. J., 34 (1985), pp. 687–704.

[14] N. J. Korevaar, *Convex solutions to nonlinear elliptic and parabolic boundary value problems*, Indiana Univ. Math. J., 32 (1983), pp. 603–614.

[15] P. Lindqvist and J. J. Manfredi, *The Harnack inequality for ∞-harmonic functions*, Electron. J. Differential Equations, 1995 (1995), pp. 1–5.

[16] S. Sakaguchi, *Concavity properties of solutions to some degenerate quasilinear elliptic Dirichlet problems*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 14 (1987), pp. 404–421.

[17] N. S. Trudinger, *On Harnack type inequalities and their applications to quasi-linear elliptic equations*, Comm. Pure Appl. Math., 20 (1967), pp. 721–747.

SIAM J. Math. Anal.
Vol. 33, No. 3, pp. 570–588

A MATHEMATICAL MODEL OF THE SPREAD OF FELINE LEUKEMIA VIRUS (FeLV) THROUGH A HIGHLY HETEROGENEOUS SPATIAL DOMAIN[*]

W. E. FITZGIBBON[†], M. LANGLAIS[‡], AND J. J. MORGAN[§]

Abstract. We are concerned with a system of partial differential equations modeling the spread of feline leukemia virus (FeLV) through highly heterogeneous habitats or spatial domains. Our differential equations may feature discontinuities in the coefficients of divergence from differential operators and discontinuities in the coupling terms. Global well posedness, long term behavior, approximation, and homogenization results are provided.

Key words. compartmental or diffractive diffusion, mass action, proportionate mixing, spatial heterogeneity, homogenization

AMS subject classifications. Primary, 92D30, 35K57; Secondary, 35R05, 35B27

PII. S0036141000371757

1. Introduction.

We shall be concerned with the development and analysis of diffusive models describing the spread of feline leukemia virus (FeLV) through the domestic cat population, *Felis catus.* The domestic cat population provides excellent examples or test cases for validating epidemic models and calibrating the effects of various factors such as spatial structure and social organization on the dynamics of the transmission of disease; see Courchamp et al. [3] and Fromont et al. [11], [12].

FeLV is an oncogenic and immunosuppressive retrovirus of the domestic cat population which can have a severe impact on population growth via direct mortality and the reduction of female fertility. An excellent review of the disease is given by Hardy [14]. Although its transmission mode may be epigenic (vertical from mother to feotus during pregnancy) (see Hoover [16]), most pregnancies abort or the newborns die quickly [16]. Consequently, we shall assume no vertical transmission. The horizontal transmission from cat to cat results from pathogens contained mainly in the saliva and occurs from biting, grooming, copulating, and sharing of food sources.

The clinical course of FeLV begins with a latent period lasting on average from three weeks to four months. Although the virus is present, this phase of the disease is asymptomatic, and because their viremia is low the cats are minimally infectious [14]. The latent phase of the disease ends with two possible outcomes. Approximately two thirds of the infected cats cease viral replication and become immune to subsequent infection. These cats are considered to be clinically recovered and appear to maintain immunity for the rest of their lives. We should, however, point out that the issue of permanent immunity is somewhat problematic and open to discussion. About one third of the infected cats become persistently viremic and develop prolif-

[*] Received by the editors May 1, 2000; accepted for publication March 28, 2001; published electronically September 5, 2001.
http://www.siam.org/journals/sima/33-3/37175.html

[†] Department of Mathematics, University of Houston, Houston, TX 77204-3476 (fitz@math.uh.edu).

[‡] UMR CNRS 5466, Mathématiques Appliquées de Bordeaux, UFR MI2S,BP 26, Université Victor Segalen Bordeaux 2, 146 rue Léo-Saignat, 33076 Bordeaux Cedex, France (langlais@mi2s.u-bordeaux2.fr).

[§] Department of Mathematics, Texas A & M University, College Station, TX 77843-3368 (jmorgan@math.tamu.edu).

erative and immunosupressive disorders which lead to death within a few weeks to several years.

Neglecting spatial considerations for the moment, we shall assume that the feline population, in the absence of infection, satisfies a basic logistic growth described by the ordinary differential equation

$$\dot{p} = (b - m)p - kp^2, \tag{1.1}$$

where the linear term accounts for the natural growth rate, b the birth rate and m the death rate, and the quadratic term accounts for the damping of growth due to resource limitations of the habitat or environment. Our basic FeLV circulation model subdivides the population into three distinct groups: susceptibles or individuals who are not infected but are capable of being infected will be denoted by u; the infective class consisting of individuals who are infected and are capable of transmitting the disease will be denoted by v; and finally, we use the variable w to denote the recovered or immune class which is made up of individuals who have been exposed to the disease but who at the conclusion of the period of recovery have gained complete immunity as a consequence of their exposure.

The following coupled system of ordinary differential equations will describe the epidemiology described heretofore:

$$\begin{aligned} du/dt &= -f(u, v, w) + b(u + w) - (m + kp)u, \\ dv/dt &= \pi f(u, v, w) - \alpha v - (m + kp)v, \\ dw/dt &= (1 - \pi)f(u, v, w) - (m + kp)w, \end{aligned} \tag{1.2}$$

with initial conditions $u(0) = u_0 > 0,\ v(0) = v_0 > 0,\ w(0) = w_0 > 0$. We set the total population $p = u + v + w$. Here some remarks are clearly in order. The linear term $b(u + w)$ represents the birth process into the susceptible class. Our model system assumes there are no births from the infective class. This represents an approximation of the underlying biological reality. We also postulate that individuals born from the recovered or immune class are born into the susceptible class and do not have immunity. This is also consistent with what actually happens: kittens born to immune females are initially passively protected by maternal antibodies but shortly become susceptible [16]. We have also collapsed the exposed class and assumed direct passage from the susceptible class to either the fully infected and infectious class or to the recovered and immune class. In all three classes we assume the logistic damping of (1.1). We account for disease-induced mortality with the term αv in the fully infected class. Last, π is the proportion of infected cats becoming infective.

The remaining term to be discussed is the incidence $f(u, v, w)$, and the rationale for the proper definition of the incidence term is an ongoing subject of discussion in the literature; see Busenberg and Cooke [2] and Diekmann et al. [4]. Two different incidence terms are commonly proposed: mass action and proportionate mixing. In the case of FeLV, field studies indicate that each of the two models can be appropriate depending on the nature of the habitat [12]: proportionate mixing appears most appropriate for rural to semiurban environments with large host population densities, and mass action is most appropriate for rural environments with smaller host population densities. We can also discuss the effects of vaccination against FeLV. However, this shall be taken up in future work treating more complete models; see also Lubkin et al. [21] and Fromont et al. [11], [12].

2. Spatially heterogeneous models. In this section we wish to include the spatial distribution of our populations and, in particular, account for the effects of spatial heterogeneity. We consider a bounded two dimensional habitat or spatial domain $\Omega \subset \mathbb{R}^2$. Because we shall not be interested in special quantitative effects contributed by boundary irregularity, we shall assume that $\partial\Omega$ is smooth with Ω lying locally on one side of $\partial\Omega$. Our state variables u, v, w no longer represent populations but rather population densities $u(x,t), v(x,t), w(x,t)$ for the susceptible, infected and infectious, and the immune/recovered populations, respectively. The time dependent populations $u(t), v(t), w(t)$ may be computed by integration over the habitat, i.e.,

$$\begin{aligned} u(t) &= \int_\Omega u(x,t)dx, \\ v(t) &= \int_\Omega v(x,t)dx, \qquad t > 0, \\ w(t) &= \int_\Omega w(x,t)dx. \end{aligned} \tag{2.1}$$

We shall account for the assumption that our populations remain confined to Ω for all time by the imposition of standard no flux boundary conditions.

The heterogeneity of the habitat will be reflected by our designation of open subregions $\{\Omega_1, \ldots, \Omega_K\} \subset \Omega$ having the properties that

$$\begin{aligned} &\overline{\Omega}_l \subset \Omega && \text{for } l = 1, \ldots, K, \\ &\overline{\Omega}_l \cap \overline{\Omega}_k = \emptyset && \text{if } l \neq k, \end{aligned} \tag{2.2}$$

and the same smoothness properties as Ω. For convenience we let

$$\Omega^* = \bigcup_{l=1}^{K} \Omega_l$$

and

$$\Omega_0 = \Omega - \overline{\Omega}^*.$$

For the purpose at hand we can consider $\Omega_1, \ldots, \Omega_K$ to be urban areas of habitat Ω with large cat population densities and the ambient region Ω_0 to be the rural part of Ω with smaller cat population densities.

The feline population will be assumed to disperse via a diffusive mechanism. However, we shall assume a sharp transition between each Ω_l and the ambient region. We model this with a compartmental or diffractive operator on Ω. A diffusion advection operator,

$$\partial u/\partial t = \operatorname{div}(d(x)\nabla u - u\bar{c}(x)),$$

is said to be compartmental or diffractive if the following conditions hold: there exist positive numbers $\underline{d}, \overline{d}$ such that

$$0 < \underline{d} \leq d(x) \leq \overline{d} \qquad \text{for } x \in \Omega, \tag{2.3}$$

and

$$d(x) = \begin{cases} d_l(x) & \text{for } x \in \Omega_l,\ l = 1 \text{ to } K, \\ d_0(x) & \text{for } x \in \Omega_0, \end{cases} \tag{2.4}$$

with $d_l(\ ) \in C(\overline{\Omega}_l)$ for $l = 1$ to K and $d_0(\ ) \in C(\overline{\Omega}_0)$. We allow for discontinuity of $d(\ )$ across the interface of Ω_l with Ω_0. We shall assume that the transport vector field $\overline{c}(\ )$ is continuous on $\overline{\Omega}$, and we impose compatibility conditions on the interface of Ω_l with Ω_0, namely, that

$$\text{(2.5)} \qquad \begin{cases} [u]_{\partial\Omega_l} = 0, \\ [d\ \partial u/\partial\eta_l]_{\partial\Omega_l} = 0, \end{cases} \qquad l = 1 \text{ to } K,$$

where $[\rho]_{\partial\Omega_l}$ denotes the saltus of the function ρ on $\partial\Omega_l$ and η_l is a unit normal vector to $\partial\Omega_l$. We point out that if u represents a concentration density, (2.5) insures continuity of the state variable u and the flux $d\ \partial u/\partial\eta_l - u\overline{c}\cdot\eta_l$ across the interface of Ω_l with Ω_0. However, a discontinuity of $d(\ )$ across $\partial\Omega_l$ will force a jump discontinuity of the normal derivative $\partial u/\partial\eta_l$ on $\partial\Omega_l$. Thus we cannot expect the smoothness from compartmental diffusive operators that we expect from normal diffusion processes. Finally, we impose boundary conditions on $\partial\Omega$; in our case we shall have homogeneous Neumann boundary conditions

$$\text{(2.6)} \qquad d(x)\partial u/\partial\eta - u\overline{c}(x)\cdot\eta = 0 \quad \text{on } \partial\Omega.$$

We point out that compartmental or diffractive diffusion operators arose in nuclear reactor engineering, and they have been studied both in Russia and the United States by Ladyzhenskaya, Rivkind, and Ural'ceva [19], Seftel [25], and Stewart [27], [28], [29]. Recently they were the subject of a throughgoing study in Horton [17]; see also Fitzgibbon and Morgan [6] and Fitzgibbon, Hollis, and Morgan [7].

We shall also want to consider reaction diffusion equations (as well as systems) with compartmental diffusion. A compartmental reaction diffusion model has the form

$$\text{(2.7)} \qquad \partial u/\partial t - \operatorname{div}(d(x)\nabla u - u\overline{c}(x)) = g(x, u)$$

for $x \in \Omega^* \bigcup \Omega_0$ and $t > 0$, where we assume that the piecewise continuity conditions on $d(\ )$, the continuity conditions on $\overline{c}(\ )$, the compatibility conditions on $\partial\Omega_l$, and the exterior homogeneous Neumann condition on $\partial\Omega$ are satisfied. We assume that $g \in C^1(\Omega_i \times \mathbb{R}_+)$ for each i, and there exists $L \in L_\infty(\Omega)$ such that L can be extended to a continuous function on each of $\overline{\Omega}_0, \overline{\Omega}_1, \ldots, \overline{\Omega}_K$ and

$$\mid g(x, u) - g(x, v) \mid \leq L(x) \mid u - v \mid, \qquad u, v \in \mathbb{R},\ x \in \Omega.$$

A strong solution to the initial value problem with $u(\ , 0) = u_0 \in C(\overline{\Omega})$ is a continuous mapping $u(\ , t)$ from $[0, \infty)$ to $L_p(\Omega)$ having the properties that $u \in W_p^{2,1}(X \times (\tau, \infty))$ for every open set Q with $\overline{Q} \subset \Omega^* \bigcup \Omega_0$ for every $\tau > 0$, and such that u satisfies the differential equation, the boundary and compatibility are continuous almost everywhere (a.e.) on Ω and $\partial\Omega$. A classical solution is one with $u(\ , t)$ as a continuous mapping of $[0, \infty)$ to $C(\overline{\Omega})$ with $u(\ ,\ ) \in C^{2,1}(\Omega_l \times (0, \infty))$ for $l = 0, 1, 2, \ldots, K$ such that the compatibility conditions (2.5) on $\partial\Omega_l$, the exterior boundary condition (2.6), and the differential equation are satisfied.

It is known that the infinitesimal generator A_p of an analytic semigroup $\{T_p(t) \mid t \geq 0\}$ on L_p can be associated with $\operatorname{div}(d(x)\nabla u - u\overline{c}(x))$ (and the given compatability conditions and boundary conditions), and it can also be shown that this differential operator can be used to obtain the infinitesimal generator A of an analytic semigroup

$\{T(t) \mid t \geq 0\}$ on $C(\overline{\Omega})$; c.f. [17], [29]. Moreover, strong L_p solutions exist and are represented by

$$u(t) = T_p(t)u(0) + \int_0^t T_p(t-s)g(\cdot, u(s))ds,\ u(0) \in L_p(\Omega), \tag{2.8}$$

and classical solutions are given by

$$u(t) = T(t)u(0) + \int_0^t T(t-s)g(\cdot, u(s))ds,\ u(0) \in C(\overline{\Omega}). \tag{2.9}$$

The regularity theory for parabolic equations with discontinuous coefficients will insure that strong $L_p(\Omega)$ solutions are regularized to classical solutions on $C(\overline{\Omega})$ (see [19]). If for each $x \in \Omega$ the forcing term $g(x,)$ is only locally Lipschitz continuous in the state variable, then we are guaranteed local classical solutions on maximal intervals $[0, T_0)$, and the key to extending these solutions will be finding continuous functions $M(t)$ on $[0, \infty)$ which bound the $L_\infty(\Omega)$ norm of the state variable $u(t)$. These results immediately extend to systems of reaction diffusion equations with compartmental diffusion.

We let $\chi(x)$ be the characteristic function of Ω^* and assume $\sigma(x)$ is a bounded strictly positive function which is piecewise continuous on $\overline{\Omega}_0, \overline{\Omega}_1, \ldots, \overline{\Omega}_K$. We obtain a heterogeneous incidence term by specifying

$$f(x, u, v, w) = \chi(x)\sigma(x)\frac{uv}{u+v+w} + (1 - \chi(x))\sigma(x)uv. \tag{2.10}$$

A moment's reflection will convince the reader that for fixed $x \in \Omega$, $f(x, , ,)$ can be made to be continuous on $\mathbb{R}^3_+$ by defining $f(x, 0, 0, 0) = 0$ and that $f(, , ,)$ is piecewise continuous on $\Omega \times \mathbb{R}^3_+$. Moreover, when $x \in \Omega^*$ we have a proportionate mixing term, and when $x \in \Omega_0$ we have mass action.

We assume that our population disperses via a compartmental diffusion system of partial differential equations of the form

(2.11)

$$\begin{aligned}
\partial u/\partial t &= \operatorname{div}(d^{(1)}(x)\nabla u - u\bar{c}^{(1)}(x)) - f(x, u, v, w) + b(x)(u + w) - (m(x) + k(x)p)u,\\
\partial v/\partial t &= \operatorname{div}(d^{(2)}(x)\nabla v - v\bar{c}^{(2)}(x)) + \pi f(x, u, v, w) - \alpha(x)v - (m(x) + k(x)p)v,\\
\partial w/\partial t &= \operatorname{div}(d^{(3)}(x)\nabla w - w\bar{c}^{(3)}(x)) + (1 - \pi)f(x, u, v, w) - (m(x) + k(x)p)w,
\end{aligned}$$

where $p = u + u + w$. We impose standard no flux boundary conditions on $\partial\Omega$,

$$\begin{aligned}
&d^{(1)}(x)\partial u/\partial\eta - u\bar{c}^1(x) \cdot \eta = 0,\\
&d^{(2)}(x)\partial v/\partial\eta - v\bar{c}^2(x) \cdot \eta = 0, \quad x \in \partial\Omega, \quad t > 0,\\
&d^{(3)}(x)\partial w/\partial\eta - w\bar{c}^3(x) \cdot \eta = 0,
\end{aligned} \tag{2.12}$$

and we impose compatibility conditions on the interface of Ω_l $(l = 1, \ldots, K)$ with Ω_0,

$$[u]_{\partial\Omega_l} = [v]_{\partial\Omega_l} = [w]_{\partial\Omega_l} = 0, \tag{2.13}$$

and

$$[d^{(1)}\partial u/\partial\eta_l]_{\partial\Omega_l} = [d^{(2)}\partial v/\partial\eta_l]_{\partial\Omega_l} = [d^{(3)}\partial w/\partial\eta_l]_{\partial\Omega_l} = 0. \tag{2.14}$$

The initial data is assumed to be continuous and nonnegative on $\overline{\Omega}$:

$$u(x, 0) = u_0(x), \quad v(x, 0) = v_0(x), \quad w(x, 0) = w_0(x). \tag{2.15}$$

Throughout what follows we shall assume that the diffusitivities $d^{(i)}(x),\ i=1,2,3,$ satisfy the conditions outlined at the beginning of the section and that the transport vector fields $\overline{c}^{(i)}(x),\ i=1,2,3,$ are continuous on $\overline{\Omega}$. The spatially dependent birth rate $b(x)$, death rate $m(x)$, logistic coefficient $k(x)$, and disease-induced mortality rate $\alpha(x)$ should all be continuous and strictly positive on $\overline{\Omega}$. The proportion π of infected cats becoming infective is constant, $0<\pi<1$ (see [11] and [21]).

In what follows we shall need to make use of b_0, b_1, k_0, k_1 such that

$$\left\{\begin{array}{l} 0<b_0<b(x)\le b_1<\infty \\ 0<k_0\le k(x)\le k_1<\infty \end{array}\right. \quad \text{for } x\in\overline{\Omega}, \tag{2.16}$$

and the following estimates.

Lemma 1. *If $u(x,t), v(x,t), w(x,t)$ are nonnegative classical solutions to* (2.11)–(2.15) *on $\Omega\times[0,T_1)$ and $p(x,t)=u(x,t)+v(x,t)+z(x,t)$, then*

$$\parallel p(\ ,t)\parallel_{1,\Omega}\le \max(\parallel p_0\parallel_{1,\Omega},(b_1/k_0)\mid\Omega\mid)=C_1; \tag{2.17}$$

if $T_1=\infty$, then

$$\lim_{t\to\infty}\sup\parallel p(\ ,t)\parallel_{1,\Omega}\le (b_1/k_0)\mid\Omega\mid. \tag{2.18}$$

Moreover, if τ,τ^* are such that $(\tau,\tau+\tau^*)\subset[0,T_1)$, then there exists a constant $C_{\tau,\tau^*}=C_{\tau,\tau^*}(b_1,k_0,\parallel p_0\parallel_{1,\Omega})$ so that for $Q(\tau,\tau+\tau^*)=\Omega\times(\tau,\tau+\tau^*)$ we have

$$\parallel p(\ ,\ )\parallel_{2,Q(\tau,\tau+\tau^*)}\le C_{\tau,\tau^*}. \tag{2.19}$$

If τ is sufficiently large, then C_{τ,τ^*} can be chosen independent of $\parallel p_0\parallel_{1,\Omega}$.

Proof. The first two estimates (2.17) and (2.18) are essentially the substance of Proposition 2 in Fitzgibbbon and Langlais [5]. To obtain the third estimate we add the components of our system to obtain

$$\frac{d}{dt}\int_\Omega p(x,t)dx\le b_1\int_\Omega p(x,t)dx-k_0\int_\Omega p^2(x,t)dx. \tag{2.20}$$

Applying the Hölder inequality and then Young's inequality to the first term on the right-hand side of (2.20), we obtain

$$\frac{d}{dt}\int_\Omega p(x,t)dx\le\tilde{b}-\tilde{k}\int_\Omega p^2(x,t)dx \tag{2.21}$$

for appropriately chosen $\tilde{b}$ and $\tilde{k}$, from which the desired result follows. □

We now turn our attention to the global well posedness and uniform boundedness of solutions to (2.11)–(2.15). We remark that if the functions d_i and c_i are in $C^1(\overline{\Omega})$, then these results can be obtained by application of the intermediate sum argument of Morgan [10], [22], [23]. However, the estimates obtained in this manner require control of the modulus of continuity of the coefficients of the differential operators, and for what follows we shall need estimates which are independent of the modulus of continuity of the coefficients. Toward this end, we have the following two theorems.

Theorem 1. *If the initial data (u_0,v_0,w_0) is nonnegative and continuous on $\overline{\Omega}$, then* (2.11)–(2.15) *have globally defined unique classical solutions $u(x,t), v(x,t), w(x,t)$ on $Q(0,\infty)=\Omega\times(0,+\infty)$ which are nonnegative.*

Proof. In light of the previous discussion concerning local well posedness, we shall be content with establishing a priori $L_\infty(\Omega)$ estimates on the state variables u, v, w. We refer the reader to Horton [17] for a careful development of the local theory. Because the vector field defined by the epidemiological kinetics and written as

$$F(x,u,v,w) = \begin{pmatrix} -f(x,u,v,w) + b(x)(u+w) - (m(x)+k(x)p)u, \\ \pi f(x,u,v,w) - \alpha(x)v - (m(x)+k(x)p)v, \\ (1-\pi)f(x,u,v,w) - (m(x)+k(x)p)w \end{pmatrix} \tag{2.22}$$

does not point out $\mathbb{R}^3_+$ on the coordinate hyperplane, $\mathbb{R}^3_+$ may be shown to be an invariant rectangle by the maximum principle arguments in [26]. To extend our local results globally we follow standard parabolic methodology and construct an $M(t) \geq 0$ which is continuous on $\mathbb{R}_+$ so that

$$\max\{\| u(\ ,t) \|_{\infty,\Omega}, \| v(\ ,t) \|_{\infty,\Omega}, \| w(\ ,t) \|_{\infty,\Omega}\} \leq M(t) \tag{2.23}$$

on the maximal interval of existence $[0, T_{\max})$. These time dependent a priori L_∞ bounds will guarantee that $T_{\max} = \infty$. We shall use energy-type arguments to bootstrap the estimates of Lemma 1 to sufficiently high L_p estimates. We will use these estimates to obtain L_p estimates of the coupling terms of (2.11). This will allow us to apply the regularity theorem of parabolic equations with discontinuous coefficients to obtain the desired function $M(t)$.

We begin with the first equation of (2.11). If we multiply by u and integrate on Ω, we obtain

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\int_\Omega u^2(x,t)dx + \int_\Omega d^{(1)}(x)|\nabla u|^2(x,t)dx + \int_\Omega k(x)u^3(x,t)dx \\ &\leq \int_\Omega b(x)u^2(x,t)dx + \int_\Omega b(x)w(x,t)u(x,t)dx + \int_\Omega u(x,t)\overline{c}^{(1)}(x)\cdot\nabla u(x,t)dx \\ &\qquad \leq b_1\left[\int_\Omega u^2(x,t)dx + \left[\int_\Omega w^2(x,t)dx\right]^{\frac{1}{2}}\left[\int_\Omega u^2(x,t)dx\right]^{\frac{1}{2}}\right] \\ &\qquad + \| \overline{c}^{(1)} \|_{\infty,\Omega}\left[\left(\int_\Omega u^2(x,t)dx\right)^{\frac{1}{2}}\left(\int_\Omega |\nabla u|^2(x,t)dx\right)^{\frac{1}{2}}\right]. \end{aligned} \tag{2.24}$$

We recall that Lemma 1 guarantees space time cylinder estimates in L_2. Integration and Young's inequality yield

$$\begin{matrix} u(\ ,t) \in L_2(\Omega) \\ u \in L_3(Q(0,T)) \\ |\nabla u| \in L_2(Q(0,T)) \end{matrix} \qquad \text{for } T < T_{\max}, \tag{2.25}$$

and the existence of functions $C_2, C_3 \in C(\mathbb{R}_+)$ so that

$$\begin{matrix} \| u(\ ,t) \|_{2,\Omega} \leq C_2(t) & \text{for } t < T_{\max}, \\ \| u \|_{3,Q(0,T)} \leq C_3(T) & \text{for } [0,T] \subset [0,T_{\max}), \\ \| |\nabla u| \|_{2,Q(\tau,T)} \leq C_3(T) & \text{for } [\tau,T] \subset [0,T_{\max}). \end{matrix} \tag{2.26}$$

We now multiply the first equation of (2.11) by u^2 and integrate on Ω to obtain

$$\frac{1}{3}\frac{d}{dt}\int_\Omega u^3(x,t)dx + 2\int_\Omega d^{(1)}(x)u(x,t)\mid\nabla u\mid^2(x,t) + \int_\Omega k(x)u^4(x,t)dx \tag{2.27}$$

$$= \int_\Omega b(x)u^3(x,t)dx + \int_\Omega b(x)w(x,t)u^2(x,t)dx + 2\int_\Omega u^2(x,t)\overline{c}^{(1)}(x).\nabla u(x,t)dx$$

$$\leq b_1\left[\int_\Omega u^3(x,t)dx + \left(\int_\Omega w^2(x,t)\right)^{\frac{1}{2}}\left(\int_\Omega u^4(x,t)\right)^{1/2}\right]$$

$$+2\parallel \overline{c}^{(1)}\parallel_{\infty,\Omega}\left[\left(\int_\Omega u^4(x,t)\right)^{1/2}\left(\int_\Omega\mid\nabla u\mid^2(x,t)dx\right)^{1/2}\right].$$

From this we may obtain

$$\begin{array}{l} u(\ ,t)\in L_3(\Omega) \text{ for } t\in[0,T_{\max}),\\ u\in L_4(Q(0,T)) \text{ for } T\in[0,T_{\max}),\end{array} \tag{2.28}$$

and $C_4, C_5 \in C(\mathbb{R}_+)$ so that

$$\begin{array}{l} \parallel u(\ ,t)\parallel_{3,\Omega}\leq C_4(t) \text{ for } t\in[0,T_{\max})\\ \parallel u\parallel_{4,Q(0,T)}\leq C_5(T) \text{ for } T\in[0,T_{\max})\end{array}, \tag{2.29}$$

We now consider the second equation of (2.11): multiply by v to obtain

$$\frac{1}{2}\frac{d}{dt}\int_\Omega v^2(x,t)dx + \int_\Omega d^{(2)}(x)\mid\nabla v\mid^2(x,t)dx + k_0\int_\Omega v^3(x,t)dx \tag{2.30}$$

$$\leq \pi\parallel\sigma\parallel_{\infty,\Omega}\left[\int_\Omega v^2(x,t)dx + \int_\Omega u(x,t)v^2(x,t)dx\right]$$

$$+\parallel\overline{c}^{(2)}\parallel_{\infty,\Omega}\left[\int_\Omega v^2(x,t)dx\right]^{1/2}\left[\int_\Omega\mid\nabla v\mid^2(x,t)dx\right]^{1/2},$$

and integrate and apply Young's inequality to obtain $d_0>0, \tilde{k}>0, \tilde{c}>0$ so that

$$\frac{1}{2}\int_\Omega v^2(x,t)dx + d_0\int_0^t\int_\Omega\mid\nabla v\mid^2(x,t)dxdt + \tilde{k}\int_0^t\int v^3(x,t)dxdt \tag{2.31}$$

$$\leq \pi\parallel\sigma\parallel_{\infty,\Omega}\left[\int_0^t\int_\Omega v^2(x,t)dxdt + \int_0^t\int_\Omega u^3(x,t)dxdt\right]$$

$$+\tilde{c}\int_0^t\int_\Omega v^2(x,t)dxdt + \frac{1}{2}\int_\Omega v_0^2(x)dx.$$

We thereby obtain time dependent estimates for $\|v(\ ,t)\|_{2,\Omega}$, $\|v\|_{3,Q(0,T)}$, and $\||\nabla v|\|_{2,Q(0,T)}$ in the same manner as these estimates were obtained for u. Estimates for $\|v(\ ,t)\|_{3,\Omega}$ and $\|v\|_{4,Q(0,T)}$ can be obtained by multiplying the second equation of (2.11) by v^2. We now multiply the third equation of (2.11) by w and use the estimates already obtained to produce estimates for $w(,t)$ in $L_3(\Omega)$ and $\mid \nabla w \mid$ in $L_2(Q(0,T))$. With these estimates in hand one proceeds to multiplying the third equation of (2.11) by w^2 to obtain estimates for $w(\ ,t)$ in $L_4(\Omega)$ and w in $L_4(Q(0,T))$.

We can now return to the first and second equations of (2.11) and multiplying by u^4 and v^4, respectively, to produce estimates on $u(\ ,t)$ in $L_5(\Omega)$ and u in $L_6(Q(0,T))$ and on $v(\ ,t)$ in $L_5(\Omega)$ and $v \in L_6(Q(0,T))$. We now have the requisite bound on u,v,w to ensure that each of three components of $F(x,u,v,w)$ is bounded in $L_3(Q(0,T))$. This allows us to apply the powerful regularity theory for parabolic equations with discontinuous coefficients in Ladyzhenskaya, Solonnikov, and Ural'ceva [20] to guarantee the existence of our desired $M(t)$ and complete our argument since $\Omega \subset \mathbb{R}^2$.

We point out that the discontinuity in the diffusitivities $d^{(i)}(x)$ prevented us from applying standard regularity theorems for reaction diffusion systems.

We find that we can also obtain uniform $L_\infty(\Omega)$ estimates.

THEOREM 2. *If* $u_0, v_0, w_0 \in C(\overline{\Omega})$ *are nonnegative and* $u(x,t), v(x,t)$*, and* $w(x,t)$ *are globally defined solutions to* (2.11)–(2.15), *then there exists a positive constant* $M_0 = M_0(u_0,v_0,w_0)$ *(independent of t) so that*

$$\max_{t>0}\{\| u(\ ,t) \|_{\infty,\Omega}, \| v(\ ,t) \|_{\infty,\Omega}, \| w(\ ,t) \|_{\infty,\Omega}\} \leq M_0.$$

Proof. If we examine the arguments of the previous global existence results and Lemma 1, we may observe that these arguments can be adapted to produce estimates for u,v,w in $L_2(Q(\tau,\tau+\tau^*))$ for any $(\tau,\tau+\tau^*) \subset [0,\infty)$ and that if τ is chosen sufficiently large, then these estimates can be chosen independent of the initial data. More specifically, we can find a uniform constant $C(2,\tau^*)$ so that

$$\max\{\| u \|_{2,Q(\tau,\tau+\tau^*)}, \| v \|_{2,Q(\tau,\tau+\tau^*)}, \| w \|_{2,Q(\tau,\tau+\tau^*)}\} \leq C(2,\tau^*).$$

Moreover, these estimates can be bootstrapped to uniform space time cylinder estimates for $u, v \in L_6(Q(\tau,\tau+\tau^*))$ and $w \in L_3(Q(\tau,\tau+\tau^*))$, i.e., we will have constants $C(6,\tau^*)$ and $C(3,\tau^*)$ so that

$$\max\{\| u \|_{6,Q(\tau,\tau+\tau^*)}, \| v \|_{6,Q(\tau,\tau+\tau^*)}\} \leq C(6,\tau^*) \tag{2.32}$$

and

$$\| w \|_{3,Q(\tau,\tau+\tau^*)} \leq C(3,\tau^*). \tag{2.33}$$

We can now use (2.10) and (2.22) to find a uniform constant $K(3,\tau^*)$ so that each component

$$\| F_i(\ ,u,v,w) \|_{3,Q(\tau,\tau+\tau^*)} \leq K(3,\tau^*) \qquad \text{for } i = 1 \text{ to } 3.$$

The regularity results of [20] for parabolic equations of the form

$$\partial z/\partial t = \operatorname{div}(d(x)\nabla z + z\overline{a}(x)) + f(x,t)$$

having discontinuous coefficients in $\Omega \subset \mathbb{R}^2$ are dependent upon the L_3 space time cylinder norm of f and the $L_\infty(\Omega)$ norm of the initial data as well as the time height of space time cylinder.

We shall solve initial value problems on short time intervals and avoid the difficulty of needing to use the $L_\infty(\Omega)$ norm of initial values using the legerdomain of constructing an auxiliary function which agrees with our unknowns on intervals but cuts off the initial value. To be more precise, we let $\varphi(t)$ be a continuously differentiable nonnegative function defined on $\mathbb{R}$ having the properties that

$$\begin{aligned} &\varphi(s) = 0 \qquad \text{for } s \leq 0,\\ &\varphi(s) = 1 \qquad \text{for } s > 1,\\ &\varphi'(s) \geq 0 \qquad \text{for } s \in (0,1). \end{aligned} \tag{2.34}$$

We shall confine our attention to the first equation. We define

$$\rho(x,t) = \varphi(t-\tau)u(x,t)$$

and observe that if $\tau \geq 0$, then the function $\rho(x,t) = u(x,t)$ for $t \in [\tau+1, \tau+2]$ and that $\rho(x,\tau) = 0$. If we differentiate ρ with respect to t, we obtain

$$\begin{aligned} \partial\rho/\partial t &= \varphi'(t-\tau)u + \operatorname{div}(d^{(1)}(x)\nabla\rho - \rho\bar{c}^{(1)}(x)) + \varphi(t-\tau)F_1(x,u,v,w)\\ &= \operatorname{div}(d^{(1)}(x)\nabla\rho - \rho\bar{c}^{(1)}(x)) + g(x,t) \end{aligned}$$

with

$$d^{(1)}(x)\partial\rho/\partial\eta - \rho c^{(1)} \cdot \eta = 0, \qquad x \in \partial\Omega, t \geq 0,$$

and

$$\rho(x,\tau) = 0.$$

Moreover, we clearly have uniform estimates for $u, v \in L_6(Q(\tau,\tau+2))$ and $w \in L_3(Q(\tau,\tau+2))$. These yield $L_\infty(Q(\tau,\tau+2))$ estimates for $\rho(x,t)$ and hence uniform $L_\infty(Q(\tau+1,\tau+2))$ estimates for $u(x,t)$. Hence, since

$$\| u \|_{\infty,Q(0,\infty)} \leq \max\left[\| u \|_{\infty,Q(0,1)}, \| u \|_{\infty,Q(1,\infty)}\right],$$

we have a uniform sup norm bound for u. Identical arguments work for v and w.

3. Long term behavior. Given the high degree of spatial heterogeneity of these systems, we cannot expect to obtain explicit statements describing the long term asymptotic behavior of solutions. Nevertheless, some results along these lines are possible. We begin by obtaining a priori gradient estimates.

LEMMA 2. *If u, v, w are classical nonnegative solutions to* (2.11)–(2.15), *then there exist a q $(2 < q < \infty)$ and a constant $C^*(q)$ so that for $t \in (\varepsilon, \infty)$*

$$\max_{t>\varepsilon}\{\| |\nabla u|(,t) \|_{q,\Omega},\ \| |\nabla v|(,t) \|_{q,\Omega},\ \| |\nabla w|(,t) \|_{q,\Omega}\} \leq C^*(q). \tag{3.1}$$

Proof. We return to (2.11) and use the uniform a priori estimates on the L_∞ norms of the state variables u, v, w together with the uniform cylinder estimates for $p(x,t)$ in $L_2(Q(\tau,\tau+\tau^*))$ to obtain uniform estimates for the nonlinear terms $F_i(x,u,v,w)$ in $L_2(Q(\tau,\tau+\tau^*))$. The application of Meyer's lemma for parabolic equations with discontinuous coefficients (see Bensoussan, Lions, and Papanicolaou [1]) will guarantee the existence of a q $(2 < q < \infty)$ so that $|\nabla u|, |\nabla v|$, and $|\nabla w|$ are uniformly bounded in $W_q^1(Q(\tau,\tau+\tau^*)$. Hence there exists a $C_q(\tau^*)$ so that

$$\max\{\| |\nabla u| \|_{q,Q(\tau,\tau+\tau^*)}, \| |\nabla v| \|_{q,Q(\tau,\tau+\tau^*)}, \| |\nabla w| \|_{q,Q(\tau,\tau+\tau^*)}\} \leq C_q(\tau^*). \tag{3.2}$$

We now differentiate our equations with respect to t. Setting $\theta = \partial u/\partial t$, $\varphi = \partial v/\partial t$, and $\psi = \partial w/\partial t$, we obtain

(3.3)

$$\begin{aligned}
\partial\theta/\partial t \; &= \operatorname{div}(d^{(1)}\nabla\theta - \theta\overline{c}^{(1)}) - \chi\sigma(1/p)(v\theta + u\varphi) + \chi\sigma(uv/p^2)(\theta + \varphi + \psi) \\
&\quad -(1-\chi)\sigma(v\theta + u\varphi) + b(\theta + \psi) - (m + kp)\theta - k(\theta + \varphi + \varphi)u, \\
\partial\varphi/\partial t \; &= \operatorname{div}(d^{(2)}\nabla\varphi - \varphi\overline{c}^{(2)}) + \pi\chi\sigma(1/p)(v\theta + u\varphi) - \pi\chi\sigma(uv/p^2)(\theta + \varphi + \psi) \\
&\quad +\pi(1-\chi)\sigma(v\theta + u\varphi) - (m + kp)\varphi - k(\theta + \varphi + \psi)v - \alpha\varphi, \\
\partial\psi/\partial t \; &= \operatorname{div}(d^{(3)}\nabla\psi - \psi\overline{c}^{(3)}) + (1-\pi)\chi\sigma(1/p)(v\theta + u\varphi) \\
&\quad -(1-\pi)\chi\sigma(uv/p^2)(\theta + \varphi + \psi)) + (1-\pi)(1-\chi)\sigma(v\theta + u\varphi) \\
&\quad -(m + kp)\psi - k(\theta + \varphi + \psi)v.
\end{aligned}$$

For notational convenience we let $\mathcal{L}_j(\rho) = \operatorname{div}(d^{(j)}\nabla\rho - \rho\overline{c}^{(j)})$. If we multiply the first equation of (2.11) through by $\partial u/\partial t$, we obtain

$$(\partial u/\partial t)^2 = \mathcal{L}_1(u)\partial u/\partial t + F_1(, u, v, w)\partial u/\partial t,$$

implying

$$(\partial u/\partial t)^2 \leq \mathcal{L}_1(u)\partial u/\partial t + K/\delta + \delta\left(\frac{\partial u}{\partial t}\right)^2 \tag{3.4}$$

for any $\delta > 0$ and appropriately chosen $K > 0$. Integration of this inequality and use of routine calculations will produce a uniform constant $C(\tau^*)$ so that for $\tau > \varepsilon$

$$\| \theta \|_{2,Q(\tau,\tau+\tau^*)} \leq C(\tau^*).$$

Analogous arguments produce uniform $L_2(Q(\tau, \tau + \tau^*))$ estimates for φ and ψ. We now return to (3.3) and observe that θ, φ, ψ satisfy

$$\begin{aligned}
\partial\theta/\partial t &= \mathcal{L}_1(\theta) + g_1(x, t), \\
\partial\varphi/\partial t &= \mathcal{L}_2(\varphi) + g_2(x, t), \\
\partial\psi/\partial t &= \mathcal{L}_3(\psi) + g_3(x, t),
\end{aligned} \tag{3.5}$$

with $g_i(,)$ uniformly bounded in $L_2(Q(\tau, \tau + \tau^*))$. We may again apply Meyer's lemma [1] to observe that φ, θ, ψ are uniformly bounded in $W_q^1(Q(\tau, \tau + \tau^*))$ for some $2 < q < \infty$. Therefore, there exists a constant $C_q(\tau^*)$ so that

$$\max\{\| \, |\nabla\theta| \, \|_{q,Q(\tau,\tau+\tau^*)}, \; \| \, |\nabla\varphi| \, \|_{q,Q(\tau,\tau+\tau^*)}, \; \| \, |\nabla\psi| \, \|_{q,Q(\tau,\tau+\tau^*)}\} \leq C_q(\tau^*). \tag{3.6}$$

We now take the minimum of the two q's for (3.2) and (3.6) and obtain uniform estimates for $|\nabla u|, |\nabla\partial u/\partial t|, |\nabla v|, |\nabla\partial v/\partial t|, |\nabla w|, |\nabla\partial w/\partial t|$ in $L_q(Q(\tau, \tau + \tau^*))$ for some $q > 2$.

It is now a simple matter to construct a priori bounds for $|\nabla u|(t)$, $|\nabla v|(t)$, and $|\nabla w|(t)$ in $L_q(\Omega)$. To see this, define

$$g(t) = \int_\Omega |\nabla u(x, t)|^q dx.$$

Then the estimates above imply $g \in W_q^1(\tau, \tau + 1)$ for every $\tau > 0$. Therefore, from the Sobolev imbedding theorem g is bounded. The same arguments can be applied to v and w. □

Each of the differential operators $\partial/\partial t - \mathcal{L}_i(\ )$, $i = 1$ to 3, can be used to define an analytic semigroup $\{T_p^i(t)/t \geq 0\}$ with infinitesimal generator A_p^i. Although the semigroup $\{T_p^i(t)/t \geq 0\}$ does not make initial data infinitely smooth, we do have the property that $T_p^i(t) : L_p(\Omega) \to D(A_p)$ for $t > 0$. Since our unknowns u, v, and w are bounded, the mappings $F_i(,)$ defining the epidemiological kinetics can be shown to be Lipschitz in $L_p(\Omega)$, and we can realize solutions to our systems as strong $L_p(\Omega)$ solutions to

$$\begin{aligned}
du/dt - A_p^1 u &= F_1(\cdot, v, v, w),\\
dv/dt - A_p^2 v &= F_2(\cdot, u, v, w),\\
dw/dt - A_p^3 w &= F_3(\cdot, u, v, w),
\end{aligned}\tag{3.7}$$

which have a variation of parameters representation:

$$\begin{aligned}
u(t) &= T_p^1(t)u_0 + \int_0^t T_p^1(t-s)F_1(\cdot, u(s), v(s), w(s))ds,\\
v(t) &= T_p^2(t)w_0 + \int_0^t T_p^2(t-s)F_2(\cdot, u(x), v(s), w(s))ds,\\
w(t) &= T_p^3(t)w_0 + \int_0^t T_p^3(t-s)F_3(\cdot, u(s), v(s), w(s))ds.
\end{aligned}\tag{3.8}$$

We can use standard semigroup theory to guarantee the local existence of nonnegative strong solutions for (3.7) if our initial data (u_0, v_0, w_0) lies in the positive cone of any $L_p(\Omega)$. We can use the regularity of the theory of parabolic equations with discontinuous coefficients [19] to argue that these local strong solutions are classical solutions which can be extended globally by use of a priori estimates.

Moreover, we can define a semidynamical system on the positive cone of any $(L_p(\Omega))^+$ by defining

$$U^p(t)[u_0, v_0, w_0]^T = [u(\ ,t), v(\ ,t), w(\ ,t)]^T,$$

where $u(\ ,t), v(\ ,t), w(\ ,t)$ are solutions to (3.7) and

$$u(x,0) = u_0(x), v(x,0) = v_0(x), w(x,0) = w_0(x).$$

We recall that a compact subset $\mathcal{A}$ is said to be a global attractor for a semidynamical system $U(t)$ defined on the positive cone X^+ of a Banach space X if $\mathcal{A}$ is forward invariant under the action of $U(t)$ and if for all $z_0 \in X^+$ we have

$$\lim_{t\to\infty} \delta(U(t)z_0, \mathcal{A}) = 0,$$

where $\delta(\ ,\ )$ is the Hausdorff metric on compact subsets of X.

We have the following result.

THEOREM 3. *If* $(u_0, v_0, w_0) \in (L_2(\Omega))^3_+$ *and* $u(\ ,t), v(\ ,t), w(\ ,t)$ *are the corresponding solutions to* (2.11)–(2.15), *then the trajectories* $\Theta = \{u(\ ,t), v(\ ,t), w(\ ,t) \mid t \geq 0\}$ *are precompact in* $L_2(\Omega)$, *and each triple of initial data* $(u_0, v_0, w_0) \in (L_2(\Omega))^3_+$ *has a compact connected forward invariant* ω*-limit set* $\omega(u_0, v_0, w_0)$. *Moreover, the semidynamical system defined as* $U(t)(u_0, v_0, w_0)^T = (u(,t), v(,t), w(,t))$ *has a global attractor* $\mathcal{A}$ *in* $(L_2(\Omega))^3_+$.

Proof. The existence of the ω-limit set is an immediate consequence of the fact that the trajectory Θ is contained in a bounded subset of $(W_q^1(\Omega))^3, q > 2$, which is compactly embedded in $(L_2(\Omega))^3$. Here we are applying standard results from the theory of semidynamical systems in functions spaces (see Hale [13]); the existence of a global attractor follows from compactness and the fact that eventually solution trajectories are uniformly bounded in $L_\infty(\Omega)$ (see Fitzgibbon, Langlais, and Morgan [8]). $\square$

4. Approximations. It is possible to approximate the compartmental model (2.11)–(2.15) by a sequence of traditional reaction diffusion models of the form

$$\begin{aligned} &\partial u_n/\partial t = \operatorname{div}(d_n^{(1)}(x)\nabla u_n - u_n\overline{c}_n^{(1)}(x)) + F_1^n(x,u_n,v_n,w_n),\\ &\partial v_n/\partial t = \operatorname{div}(d_n^{(2)}(x)\nabla v_n - v_n\overline{c}_n^{(2)}(x)) + F_2^n(x,u_n,v_n,w_n), \qquad x\in\Omega, t>0,\\ &\partial w_n/\partial t = \operatorname{div}(d_n^{(3)}(x)\nabla w_n - w_n\overline{c}_n^{(3)}(x)) + F_3^n(x,u_n,v_n,w_n), \end{aligned} \tag{4.1}$$

with homogeneous Neumann boundary conditions

$$\begin{aligned} &d_n^{(1)}\partial u_n/\partial\eta - u_n\overline{c}_n^{(1)}\cdot\eta = 0,\\ &d_n^{(2)}\partial v_n/\partial\eta - v_n\overline{c}_n^{(2)}\cdot\eta = 0, \quad x\in\partial\Omega, t>0,\\ &d_n^{(3)}\partial w_n/\partial\eta - w_n\overline{c}_n^{(3)}\cdot\eta = 0, \end{aligned} \tag{4.2}$$

and initial conditions (nonnegative and continuous on $\overline{\Omega}$)

$$u_n(x,0) = u_0(x), \quad v_n(x,0) = v_0(x), \quad w_n(x,0) = w_0(x), \quad x\in\Omega. \tag{4.3}$$

The diffusitivities $d_n^{(i)}(x)$ are assumed to be continuous on $\overline{\Omega}$ with

$$0 < \underline{d} \le d_n^{(i)}(x) \le \overline{d} \ \text{ for } i = 1 \text{ to } 3,\ x\in\Omega,\ n\in\mathbb{Z}^+,$$

and we assume that advection fields $\overline{c}_n^{(i)}(x)$ are continuous on $\overline{\Omega}$. The kinetic terms have the form

$$F^n(x,u,v,w) = \begin{pmatrix} -f_n(x,u,v,w) + b(x)(u+w) - (m(x)+k(x)p)u,\\ \pi f_n(x,u,v,w) - \alpha(x)v - (m(x)+k(x)p)v,\\ (1-\pi)f_n(x,u,v,w) - (m(x)+k(x)p)w \end{pmatrix} \tag{4.4}$$

with incidence term

$$f_n(x,u,v,w) = \chi_n(x)\sigma_n(x)uv/(u+v+w) + (1-\chi_n(x))\sigma_n(x)uv, \tag{4.5}$$

where $\chi_n(\ )$ and $\sigma_n(\ )$ belong to $C(\overline{\Omega})$, $\sigma_n(\ )$ being strictly positive and $0\le\chi_n(\ )\le 1$ on $\overline{\Omega}$.

We shall assume the following:

(A$_1$) The approximating terms $d_n^{(i)}(\ )$, $\chi_n(\ )$, and $\sigma_n(\ )$ are uniformly bounded in $L_\infty(\Omega)$;

(A$_2$) $\lim\limits_{n\to\infty} d_n^{(i)}(x) = d^{(i)}(x)$ for a.e. $x\in\Omega$, $i = 1$ to 3,

$\lim\limits_{n\to\infty} \overline{c}_n^{(i)} = \overline{c}^{(i)}$ in $C(\overline{\Omega})$, $i = 1$ to 3,

$\lim\limits_{n\to\infty} \sigma_n(x) = \sigma(x)$ for a.e. $x\in\Omega$,

$\lim\limits_{n\to\infty} \chi_n(x) = \chi(x)$ for a.e. $x\in\Omega$.

The elliptic operators $\mathrm{div}(d_n^{(i)}(x)\nabla z - z\overline{c}_n^{(i)}(x))$ can be used to specify infinitesimal generators A_n^i of analytic semigroups $\{T_n^i(t) \mid t \geq 0\}$ in the spaces $L_p(\Omega)$ $(p > 1)$ and $C(\overline{\Omega})$. Here we shall with some abuse of notation suppress the particular function space dependence of the generators and semigroups. Solutions to (4.1)–(4.3) may be represented as

$$\begin{aligned} u_n(\cdot,t) &= T_n^1(t)u_0 + \int_0^t T_n^1(t-s)F_n^1(\cdot,u_n(,s),v_n(,s),w_n(,s))ds,\\ v_n(\cdot,t) &= T_n^2(t)v_0 + \int_0^t T_n^2(t-s)F_n^2(\cdot,u_n(,s),v_n(,s),w_n(,s))ds, \qquad (4.6)\\ w_n(\cdot,t) &= T_n^3(t)w_0 + \int_0^t T_n^3(t-s)F_n^3(\cdot,u_n(,s),v_n(,s),w_n(,s))ds. \end{aligned}$$

We have pointed out that solutions to (2.11)–(2.15) can be represented by the Duhamel formulae

$$\begin{aligned} u(\cdot,t) &= T^1(t)u_0 + \int_0^t T^1(t-s)F^1(\cdot,u(,s),v(,s),w(,s))ds,\\ v(\cdot,t) &= T^2(t)v_0 + \int_0^t T^2(t-s)F^2(\cdot,u(,s),v(,s),w(,s))ds, \qquad (4.7)\\ w(\cdot,t) &= T^3(t)w_0 + \int_0^t T^3(t-s)F^3(\cdot,u(,s),v(,s),w(,s))ds. \end{aligned}$$

We have the following approximation theorem.

Theorem 4. *If the approximation conditions* (A_1) *and* (A_2) *hold and the solutions to* (4.1)–(4.3) *are given by* $u_n(x,t)$, $v_n(x,t)$, *and* $w_n(x,t)$ *and the solution to the compartmental system is given by* $u(x,t)$, $v(x,t)$, $w(x,t)$, *then on any interval* $[0,T]$

$$\begin{aligned} &\lim_{n\to\infty} \| u_n(\cdot,t) - u(\cdot,t) \|_{2,\Omega} = 0,\\ &\lim_{n\to\infty} \| v_n(\cdot,t) - v(\cdot,t) \|_{2,\Omega} = 0,\\ &\lim_{n\to\infty} \| w_n(\cdot,t) - w(\cdot,t) \|_{2,\Omega} = 0 \qquad \textit{for } t \in [0,T]. \end{aligned}$$

Proof. We shall establish subsequence convergence which will show that any sequence has a subsequence converging to a solution of (2.11)–(2.15). However, uniqueness of the limit guarantees that the subsequential convergence is in fact convergence. If we return to estimates producing Lemma 1, we see that our estimates guarantee a priori $L_\infty(\Omega)$ bounds and $L_2(\Omega)$ gradient bounds which depend only upon the size of the diffusion coefficients and the initial data, and the bounds for the birth rate and logistic coefficients.

Therefore, we can obtain uniform bounds for $\| |\nabla u_n|(,t) \|_{2,\Omega}, \| |\nabla v_n|(,t) \|_{2,\Omega}$, and $\| |\nabla w_n|(,t) \|_{2,\Omega}$, and hence for each $t \in [0,T]$ the solutions lie in a bounded subset of $H^1(\Omega)$. Because $H^1(\Omega)$ is compactly embedded in $L_2(\Omega)$, a standard application of the Arzela–Ascoli lemma will guarantee the convergence of a subsequence $u_{n'}(,t), v_{n'}(,t), w_{n'}(,t)$ in $L_2(\Omega)$ to a limiting function $(u^*(\ ,t), v^*(\ ,t), w^*(\ ,t))$. This subsequence has a subsequence $(u_{n''}(\ ,t), v_{n''}(\ ,t), w_{n''}(\ ,t))$ which converges pointwise a.e. to $(u^*(\ ,t), v^*(\ ,t), w^*(\ ,t))$. Standard semigroup approximation theory (see Pazy [24]) will insure the strong convergence of $T_n^1(t), T_n^2(t)$, and $T_n^3(t)$ to $T^1(t), T^2(t)$, and $T^3(t)$.

Moreover, we have pointwise a.e. convergence of $F_n^i(x, u_n(x,t), v_n(x,t), w_n(x,t))$ to the corresponding $F^i(x, u^*(x,t), v^*(x,t), w^*(x,t))$, and we can apply the Lebesgue convergence theorem to ensure convergence of $(u_{n''}, v_{n''}, w_{n''})$ to a solution of (4.7). The regularity theory of analytic semigroups guarantees that a solution to (4.7) is indeed a solution to (2.11)–(2.15). Because these solutions are unique, our proof is complete. □

5. Complex habitats with repeated microstructure. A different approximation is realized by averaging the effects of the complex local structure over the whole domain; by virtue of this approximation endeavor we obtain a global picture of the dynamics. More precisely, many habitats consist of more or less repeatedly interspersed fragmented subpatches. Here we shall attempt to isolate a subpatch $\Omega_\#$ of Ω and consider the fragmentation of Ω as being produced by a periodic reproduction of $\Omega_\#$. Although the scale of the microstructure of $\Omega_\#$ may be small in comparison with the scale of Ω, this microstructure can have a profound effect upon the global dynamics. Theoretically this situation can be described by the methods presented heretofore. However, from a practical computational point of view, the microstructure of the domain may be far too fine to track, and it becomes reasonable to implement homogenization techniques which average the effects of the local structure and predict the overall dynamics.

We commence with a mathematical formulation. We introduce a basic cell

$$\hat{\Omega} = \prod_{j=1}^{2} [0, Y_j^0] \subset \mathbb{R}^2. \tag{5.1}$$

A function $\varphi(\,) : \mathbb{R}^2 \to \mathbb{R}$ is said to be $\hat{\Omega}$ periodic if it admits a period Y_j^0 in the direction y_j, $j = 1, 2$. We let $d^{(1)}(y), d^{(2)}(y), d^{(3)}(y)$, $\chi(y), \sigma(y), \alpha(y), b(y), m(y), k(y)$ be functions satisfying the conditions of section 2 on $\hat{\Omega}$ with corresponding subregions $\hat{\Omega}_0, \hat{\Omega}_1, \ldots, \hat{\Omega}_K$. We extend $d^{(1)}(\,), d^{(2)}(\,), d^{(3)}(\,), \chi(\,), \sigma(\,), \alpha(\,), b(\,), m(\,), k(\,)$ periodically to $\mathbb{R}^2$ with the distinguished subregions being reproduced as well. We define

$$\begin{aligned} &d_\varepsilon^{(i)}(x) = d^{(i)}(x/\varepsilon), \\ &\overline{c}_\varepsilon^{(i)}(x) = \overline{c}^{(i)}(x/\varepsilon) \qquad \text{for } i = 1 \text{ to } 3,\ x \in \Omega, \end{aligned} \tag{5.2}$$

and

$$\begin{aligned} &\chi_\varepsilon(x) = \chi(x/\varepsilon), \quad \sigma_\varepsilon(x) = \sigma(x/\varepsilon), \\ &b_\varepsilon(x) = b(x/\varepsilon), \quad m_\varepsilon(x) = m(x/\varepsilon),\ k_\varepsilon(x) = k(x/\varepsilon), \\ &\alpha_\varepsilon(x) = \alpha(x/\varepsilon), \quad f_\varepsilon(x,u,v,w) = f(x/\varepsilon, u, v, w). \end{aligned} \tag{5.3}$$

In the same manner we get corresponding subregions $\Omega_{0,\varepsilon}, \Omega_{1,\varepsilon}, \ldots, \Omega_{K,\varepsilon}$.

For small ε we consider the system of partial differential equations

(5.4)

$$\begin{aligned} \partial u_\varepsilon/\partial t &= \operatorname{div}(d_\varepsilon^{(1)} \nabla u_\varepsilon - u_\varepsilon \overline{c}_\varepsilon^{(1)}) - f_\varepsilon(u_\varepsilon, v_\varepsilon, w_\varepsilon) + b_\varepsilon(u_\varepsilon + w_\varepsilon) - (m_\varepsilon + k_\varepsilon p_\varepsilon) u_\varepsilon, \\ \partial v_\varepsilon/\partial t &= \operatorname{div}(d_\varepsilon^{(2)} \nabla v_\varepsilon - v_\varepsilon \overline{c}_\varepsilon^{(2)}) + \pi f_\varepsilon(u_\varepsilon, v_\varepsilon, w_\varepsilon) - \alpha_\varepsilon v_\varepsilon - (m_\varepsilon + k_\varepsilon p_\varepsilon) v_\varepsilon, \\ \partial w_\varepsilon/\partial t &= \operatorname{div}(d_\varepsilon^{(3)} \nabla w_\varepsilon - w_\varepsilon \overline{c}_\varepsilon^{(3)}) + (1-\pi) f_\varepsilon(u_\varepsilon, v_\varepsilon, w_\varepsilon) - (m_\varepsilon + k_\varepsilon p_\varepsilon) w_\varepsilon \end{aligned}$$

with no flux boundary conditions

$$\begin{aligned} &d_\varepsilon^{(1)}\partial u_\varepsilon/\partial\eta - u_\varepsilon \bar{c}_\varepsilon^{(1)}\cdot\eta = 0,\\ &d_\varepsilon^{(2)}\partial v_\varepsilon/\partial\eta - v_\varepsilon \bar{c}_\varepsilon^{(2)}\cdot\eta = 0, \qquad x\in\partial\Omega,\\ &d_\varepsilon^{(3)}\partial w_\varepsilon/\partial\eta - w_\varepsilon \bar{c}_\varepsilon^{(3)}\cdot\eta = 0, \end{aligned} \tag{5.5}$$

and compatibility conditions

$$[u_\varepsilon]_{\partial\Omega_{l,\varepsilon}} = [v_\varepsilon]_{\partial\Omega_{l,\varepsilon}} = [w_\varepsilon]_{\partial\Omega_{l,\varepsilon}} = 0, \qquad l = 1,\dots,K, \tag{5.6}$$

$$[d_\varepsilon^{(1)}\partial u_\varepsilon/\partial\eta]_{\partial\Omega_{l,\varepsilon}} = [d_\varepsilon^{(2)}\partial v_\varepsilon/\partial\eta]_{\partial\Omega_{l,\varepsilon}} = [d_\varepsilon^{(3)}\partial w_\varepsilon/\partial\eta]_{\partial\Omega_{l,\varepsilon}} = 0$$

with $\partial\Omega_{l,\varepsilon}$ the interface between $\Omega_{l,\varepsilon}$ and $\Omega_{0,\varepsilon}$, $l = 1,\dots,K$.

We have initial conditions

$$u_\varepsilon(x,0) = u_0(x), \quad v_\varepsilon(x,0) = v_0(x), \quad w_\varepsilon(x,0) = w_0(x) \qquad \text{for } x\in\Omega. \tag{5.7}$$

A simple prototype of this scenario has been considered in Fitzgibbon, Langlais, and Morgan [9] and Heiser, Langlais, and Pontier [15]. We are guaranteed the existence of solutions to (5.4)–(5.7) by our existing theory, and our present concern shall be the behavior as $\varepsilon \downarrow 0$. A partial answer to this question is given by the following result.

Theorem 5. *For each $\varepsilon > 0$ and $T > 0$ there exists a globally defined unique classical solution triple to (5.4)–(5.7), $u_\varepsilon(x,t), v_\varepsilon(x,t), w_\varepsilon(x,t)$. There exist positive definite symmetric constant matrices $D_h^{(j)}$ and constant vector fields $\bar{c}_h^{(j)}$ depending solely on $d^{(j)}, \bar{c}^{(j)}, \Omega$, and $\hat{\Omega}$ for $j = 1$ to 3, constants*

$$\sigma = \frac{1}{|\,\hat{\Omega}\,|}\int_{\hat{\Omega}} \sigma(y)dy, \qquad \chi^\sharp = \frac{1}{\sigma}\frac{1}{|\,\hat{\Omega}\,|}\int_{\hat{\Omega}} \chi(y)\sigma(y)dy,$$

$$b = \frac{1}{|\,\hat{\Omega}\,|}\int_{\hat{\Omega}} b(y)dy, \qquad m = \frac{1}{|\,\hat{\Omega}\,|}\int_{\hat{\Omega}} m(y)dy, \qquad k = \frac{1}{|\,\hat{\Omega}\,|}\int_{\hat{\Omega}} k(y)dy,$$

and a function

$$f(u,v,w) = \chi^\sharp\sigma\frac{uv}{u+v+w} + (1-\chi^\sharp)\sigma uv$$

such that, as $\varepsilon \to 0$, $u_\varepsilon(\,,t), v_\varepsilon(\,,t), w_\varepsilon(\,,t)$ converges strongly in $L_2(0,T)$ to the classical solution $u(\,,t), v(\,,t), w(\,,t)$ of the homogenized system

$$\begin{aligned} &\partial u/\partial t = \operatorname{div}(D_h^{(1)}\nabla u - u\bar{c}_h^{(1)}) - f(u,v,w) + b(u+w) - (m+kp)u,\\ &\partial v/\partial t = \operatorname{div}(D_h^{(2)}\nabla v - v\bar{c}_h^{(2)}) + \pi f(u,v,w) - \alpha v - (m+kp)v,\\ &\partial w/\partial t = \operatorname{div}(D_h^{(3)}\nabla w - w\bar{c}_h^{(3)}) + (1-\pi) f(u,v,w) - (m+kp)w, \end{aligned} \tag{5.8}$$

with no flux boundary conditions

$$\begin{aligned} &\sum_{i,j=1}^{2} D_{h,ij}^{(1)}\frac{\partial u}{\partial x_i}\cos(\eta,x_j) - uc_h^{(1)}\cdot\eta = 0,\\ &\sum_{i,j=1}^{2} D_{h,ij}^{(2)}\frac{\partial v}{\partial x_i}\cos(\eta,x_j) - vc_h^{(2)}\cdot\eta = 0, \qquad x\in\partial\Omega, \quad t>0,\\ &\sum_{i,j=1}^{2} D_{h,ij}^{(3)}\frac{\partial w}{\partial x_i}\cos(\eta,x_j) - wc_h^{(3)}\cdot\eta = 0, \end{aligned} \tag{5.9}$$

and initial conditions

$$u(x,0)=u_0(x),\ v(x,0)=v_0(x),\ w(x,0)=w_0(x)\ \ \textit{for } x\in\Omega.$$

Proof. This result is obtained via applications of standard techniques developed in, e.g., Bensoussan, Lions, and Papanicolaou [1] and Jikov, Kozlov, and Oleinik [18]. We point out that standard arguments will yield the existence and uniqueness of solutions of the homogenized problem (c.f. Theorem 1).

The crucial step lies in using the result in Theorem 2 stating that

$$\max_{t>0}\{\parallel u_\varepsilon(,t)\parallel_{\infty,\Omega},\parallel v_\varepsilon(,t)\parallel_{\infty,\Omega},\parallel w_\varepsilon(,t)\parallel_{\infty,\Omega}\}\le M_0,$$

M_0 being a constant independent of ε for $0<\varepsilon\le 1$. Then it follows from integration by parts that for any fixed $T>0$ the solution triples $(u_\varepsilon,v_\varepsilon,w_\varepsilon)$ are bounded in the Sobolev space of order one $H^1(\Omega\times(0,T))$, independently of ε for $0<\varepsilon\le 1$. A compactness argument yields that they lie in a relatively compact subset of $L_2(\Omega\times(0,T))$. Hence there exists a sequence $(u_{\varepsilon'},v_{\varepsilon'},w_{\varepsilon'})$ converging to some limit (u,v,w) strongly in $L_2(\Omega\times(0,T))$ and weakly in $H^1(\Omega\times(0,T))$ as $\varepsilon'\to 0$.

Next it is well known that for any function $\varphi(\,):\mathbb{R}^2\to\mathbb{R}$ and $\hat\Omega$ periodic

$$\varphi_\varepsilon(\,)\rightharpoonup\mathcal{M}(\varphi)=\frac{1}{\mid\hat\Omega\mid}\int_{\hat\Omega}\varphi(y)dy,\quad\text{as }\varepsilon\to 0,$$

in a weak-star $L^\infty(\Omega)$ sense, i.e., for any $\psi\in L^1(\Omega)$

$$\int_\Omega\varphi_\varepsilon(x)\psi(x)dx\to\mathcal{M}(\varphi)\int_\Omega\psi(x)dx,\quad\text{as }\varepsilon\to 0.$$

From these two facts the convergence as $\varepsilon'\to 0$ of the kinetics on the right-hand sides of (5.4) toward the corresponding kinetics of (5.8) follows, weakly in $L_2(\Omega\times(0,T))$ and strongly in the dual space $[H^1(\Omega\times(0,T))]'$.

The goal of homogenization techniques is to handle the behavior of such quantities as $-\text{div}(d_\varepsilon^{(1)}\nabla u_\varepsilon-u_\varepsilon\overline{c}_\varepsilon^{(1)})$ when $\varepsilon\to 0$. Then one can show that there exist a positive definite symmetric matrix $D_h^{(1)}$ and a vector $\overline{c}_h^{(1)}$ depending solely on $d^{(1)},\overline{c}^{(1)},\Omega$, and $\hat\Omega$ such that upon extracting further subsequences

$$d_{\varepsilon''}^{(1)}\nabla u_{\varepsilon''}-u_{\varepsilon''}\overline{c}_{\varepsilon''}^{(1)}\rightharpoonup D_h^{(1)}\nabla u-u\overline{c}_h^{(1)}$$

weakly in $L_2(\Omega\times(0,T))$ as ε" $\to 0$; see [1] and [18]. Identical arguments work for the equations for v_ε and w_ε.

At this point the convergence of a suitable subsequence of $(u_\varepsilon,v_\varepsilon,w_\varepsilon)_{0<\varepsilon\le 1}$ toward a solution of (5.8) is established. A uniqueness argument ensures that this subsequential convergence is indeed convergence. □

Homogenized operators. We point out that computation of the homogenized diffusion operators is complicated. However, an algorithm using asymptotic expansions appears in [1] and [18].

In the one dimensional case some simplifications occur, and one has

$$D_h^{(j)}=\left[\mathcal{M}\left(\frac{1}{d^{(j)}}\right)\right]^{-1}\quad\text{and}\quad\overline{c}_h^{(j)}=D_h^{(j)}\left[\mathcal{M}\left(\frac{\overline{c}^{(j)}}{d^{(j)}}\right)\right],\ j=1\cdots 3.$$

In the two dimensional case one must first compute a set of auxiliary functions, namely, $\chi_l^{(j)}$ and $\xi^{(j)}$, $\hat{\Omega}$ periodic members of $H^1(\hat{\Omega})$ and solutions of

$$\begin{aligned} &-\operatorname{div}(d^{(j)}(y)\nabla\chi_l^{(j)}) = \frac{\partial d^{(j)}(y)}{\partial y_l},\\ &-\operatorname{div}(d^{(j)}(y)\nabla\xi^{(j)}) = -\operatorname{div}(\overline{c}^{(j)}), \qquad j = 1\cdots 3, \quad l = 1, 2. \end{aligned}$$

These nine functions are uniquely defined up to a constant; note that $\xi^{(j)}$ is a constant when $\operatorname{div}(\overline{c}^{(j)}) = 0$. The entries of the three matrices $D_h^{(j)}$ are

$$D_{h,il}^{(j)} = \begin{cases} \mathcal{M}(d^{(j)}) + \mathcal{M}\left(d^{(j)}\dfrac{\partial\chi_i^{(j)}}{\partial y_i}\right) & \text{for } l = i,\\ \mathcal{M}\left(d^{(j)}\dfrac{\partial\chi_l^{(j)}}{\partial y_i}\right) & \text{for } l \neq i, \end{cases}$$

for $j = 1$ to 3; hence these matrices may not be diagonal. The entries of the three constant vectors $c_h^{(j)}$ are

$$\overline{c}_{h,i}^{(j)} = \mathcal{M}(\overline{c}_i^{(j)}) - \mathcal{M}\left(d^{(j)}\frac{\partial\xi^{(j)}}{\partial y_i}\right), \ i = 1, 2, \quad j = 1\cdots 3.$$

Note that $\overline{c}_h^{(j)} = \mathcal{M}(\overline{c}^{(j)})$ when $\operatorname{div}(\overline{c}^{(j)}) = 0$.

6. Concluding consideration. We feel that the notions of compartmental diffusion and homogenization are important for the spatial spread of infectious disease within complex highly spatially heterogeneous habitats. Heretofore, standard Fickian diffusion has been used to describe the space time evolution or progation within homogeneous habitats supporting spatially distributed habitats. Indeed, one can incorporate heterogeneity in diffusion and transport terms; however, one expects diffusion smoothing to regularize these effects. We also feel that diffractive or compartmental diffusion offers many other interesting applications within the reaction diffusion context.

REFERENCES

[1] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North Holland, Amsterdam, 1978.

[2] S. Busenberg and K. C. Cooke, *Vertically Transmitted Diseases*, Biomathematics 23, Springer-Verlag, New York, 1993.

[3] F. Courchamp, D. Pontier, M. Langlais, and M. Artois, *Population dynamics of feline immunodeficiency virus within populations of cats*, J. Theoretical Biology, 175 (1995), pp. 553–560.

[4] O. Diekmann, M. C. M. De Jong, A. A De Koeijer, and P. Reijnders, *The force of infection in populations of varying size: A modelling problem*, J. Biological Systems, 3 (1995), pp. 519–529.

[5] W. Fitzgibbon and M. Langlais, *Diffusive SEIR models with logistic population control*, Comm. Appl. Nonlinear Anal., 4 (1997), pp. 1–16.

[6] W. Fitzgibbon and J. J. Morgan, *Diffractive diffusion systems with locally defined reactions*, in Evolution Equations, G. Goldstein et al., eds., M. Dekker, New York, 1994, pp. 177–186.

[7] W. Fitzgibbon, S. Hollis, and J. Morgan, *Steady state solutions for balanced reaction diffusion systems on heterogeneous domains*, Differential Integral Equations, 12 (1999), pp. 637–660.

[8] W. Fitzgibbon, M. Langlais, and J. J. Morgan, *Eventually uniform bounds for a quasipositive reaction diffusion system*, Japan J. Industrial Appl. Math., 16 (2000), pp. 225–241.

[9] W. Fitzgibbon, M. Langlais, and J. Morgan, *Epidemic Models with Compartmental Diffusion*, preprint, University of Houston, Houston, TX, 1999.

[10] W. Fitzgibbon, J. J. Morgan, and R. Sanders, *Global existence and boundedness for a class of inhomogeneous semilinear parabolic equations*, Nonlinear Anal., 19 (1992), pp. 885–899.

[11] E. Fromont, M. Artois, M. Langlais, F. Courchamp, and D. Pontier, *Modelling the feline leukemia virus (FeLV) in natural populations of cats*, Theoretical Population Biology, 52 (1997), pp. 60–70.

[12] E. Fromont, M. Langlais, and D. Pontier, *Dynamics of a feline retrovirus (FeLV) in host populations with variable structures*, Proc. Roy. Soc. London Ser. B, 265 (1998), pp. 1097–1104.

[13] J. Hale, *Asymptotic Behavior of Dissipative Systems*, AMS, Providence, RI, 1988.

[14] W. D. Hardy, Jr., *The virology, immunology and epidemiology of the feline leukemia virus*, in Feline Leukemia Virus, W. D. Hardy, M. Essex, and A. J. McClelland, eds., Elsevier North Holland, New York, 1980, pp. 33–78.

[15] F. Heiser, M. Langlais, and D. Pontier, *Modelling the Propagation of a Feline Retrovirus within a Heterogeneous Host Population*, preprint, Université de Bordeaux II, Bordeaux, France, 1999.

[16] E. A. Hoover, J. L. Rojko, and S. L. Quackenbush, *Congenital feline leukemia virus*, Leukemia Rev. Int., 1 (1983), pp. 7–8.

[17] P. Horton, *Global Existence of Solutions to Reaction Diffusion Systems Heterogeneous Domains*, Dissertation, Texas A & M University, College Station, TX, 1998.

[18] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, New York, 1994.

[19] O. A. Ladyzhenskaya, V. Rivkind, and N. Ural'ceva, *On the classical solvability of diffraction problems for equations of elliptic and parabolic type*, Soviet Math. Dokl., 5 (1965), pp. 1249–1252.

[20] O. A. Ladyzhenskaya, V. Solonnikov, and N. Ural'ceva, *Linear and quasilinear equations of parabolic type*, Transl. AMS 23, Providence Rhode Island, 1968.

[21] S. R. Lubkin, J. Romatowski, M. Zhu, P. M. Kulesa, and K. A. J. White, *Evaluation of feline leukemia virus control measures*, J. Theoretcial Biology, 178 (1996), pp. 53–60.

[22] J. J. Morgan, *Global existence for semilinear parabolic systems*, SIAM J. Math. Anal., 20 (1989), pp. 1128–1144.

[23] J. J. Morgan, *Boundedness and decay results for reaction-diffusion systems*, SIAM J. Math. Anal., 21 (1990), pp. 1172–1189.

[24] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, 1983.

[25] Z. Seftel, *Estimates in L_q of solutions of elliptic equations with discontinuous coefficients and satisfying general boundary conditions and conjugacy conditions*, Soviet Math. Dokl., 4 (1963), pp. 321–324.

[26] J. Smoller, *Shock Waves and Reaction Diffusion Equations*, Springer-Verlag New York, 1983.

[27] H. Stewart, *Generation of analytic semigroups by strongly elliptic operators*, Trans. Amer. Math. Soc., 199 (1974), pp. 141–162.

[28] H. Stewart, *Spectral theory of heterogeneous diffusion systems*, J. Math. Anal. Appl., 54 (1976), pp. 59–78.

[29] H. Stewart, *Generation of analytic semigroups by strongly elliptic operators under general boundary conditions*, Trans. Amer. Math. Soc., 259 (1980), pp. 299–310.

SIAM J. MATH. ANAL.
Vol. 33, No. 3, pp. 589–606

THE NISHIURA–OHNISHI FREE BOUNDARY PROBLEM IN THE 1D CASE*

PAUL C. FIFE† AND DANIELLE HILHORST‡

Abstract. A free boundary problem due to Nishiura and Ohnishi is solved in one space dimension. That problem was derived, during their study of phase separation phenomena in diblock copolymers, as an asymptotic limit of pattern-forming PDEs generalizing that of Cahn and Hilliard. The free boundary problem in one dimension reduces to a linear system of ODEs for the lengths of the intervals between interfaces. This system also arises in a completely different context as the spatial discretization of a simple heat equation in a medium with periodic properties. (The medium is homogeneous in an important special case.) The initial-value problem for this system is completely solved, and global stability results for stationary solutions (in which the interfaces are regularly spaced) are obtained. Nucleation phenomena are briefly discussed.

Key words. discrete heat equation, free boundary problem, diblock copolymers, global stability

AMS subject classifications. 76R99, 39A11

PII. S0036141000372507

1. Introduction.

Pattern-forming phenomena in block copolymer melts have been the subject of a number of field-theoretic models and analyses. Equilibrium models based on a free energy functional were given by Leibler [4], Ohta and Kawasaki [7], and Kawasaki, Ohta, and Kohrogui [3]. Bahiana and Oono [1] suggested a phenomenological cell dynamical system leading to an evolutionary PDE for the composition $u(x,t)$ of the melt as a function of space and time. In [5], a similar PDE was proposed by Nishiura and Ohnishi as a gradient (steepest descent) flow for a free energy functional $E^\epsilon[u]$, generalizing that of Ginzburg and Landau, of the type appearing in [7, 3]. This PDE was investigated in [5, 6].

In [5] the authors also derived formally a limit free boundary problem (FBP) as a parameter ϵ in the equation approaches zero. This limiting process was investigated rigorously by Henry [2] for radial solutions in three dimensions. The local and global minimizers of E^ϵ and of related functionals in one space dimension were considered by Ren and Wei [9, 10] and by Ohnishi et al. [8]. In [9], the local minimizers for small ϵ were found to be close, in the L^2 sense, to piecewise constant functions whose intervals of constancy alternated in length. Among them, the global minimizers were specified. In [8, 10], the global minimizers of a rescaled free energy functional were characterized.

In this paper we concentrate on the evolution the FBP and show that in one dimension it reduces to a linear homogeneous system of ODEs of a particularly simple type. The variables in this system are the interval lengths between the discrete interfaces. The system is the same as that which arises as spatial discretization of a heat equation. The role of the space variable in the heat equation is taken, in

*Received by the editors May 17, 2000; accepted for publication April 23, 2001; published electronically September 5, 2001.
http://www.siam.org/journals/sima/33-3/37250.html

† Mathematics Department, University of Utah, 155 South 1400 East, Salt Lake City, UT 84112-0090 (fife@math.utah.edu). This author's work was supported by NSF grant DMS 9703483 and by the University of Paris-Sud during two visits.

‡Laboratoire de Mathématiques, Bâtiment 425, Université de Paris-Sud, 91405 Orsay Cedex, France (Danielle.HILHORST@math.u-psud.fr). This author's work was supported by the TMR project Non Linear Parabolic Equations (ERB FMRX CT98 0201).

the context of the FBP, by the index ordering the positions of the intervals. Thus although the FBP is in general difficult and nonlinear, in one dimension it reduces in a remarkable way to a linear spatially discrete heat equation. This suggests that the FBP provides a very efficient way to "uniformize" a given set $\{x_n\}$ of interfaces. This point is discussed at the end of section 3.

The conversion of the FBP to a system of ODEs is given in section 3 after background considerations in section 2. The same problem for a finite interval can be reformulated as a periodic problem on the whole line. The problem on the whole real line with an infinite number of interfaces, periodic or not, is solved explicitly in section 4. In the periodic (i.e., finite interval) case the solution simplifies somewhat; this is shown in section 4.7. Since we wish the interval lengths to be positive, it is important to know whether any can collapse to zero, and it is shown in section 5 that collapsing is not possible.

All stationary solutions are found in section 6, where it is also shown how to construct certain other similarity solutions. Of special interest is a one-parameter family of stationary solutions with positive bounded interval lengths for each given average value $\bar{u}$ of the concentration variable. These solutions are those which were identified in [9] as approximating the local minimizers of the free energy when ϵ is small. The stability of stationary solutions is addressed in sections 7 (for the whole line) and 8 (for a bounded interval). In the latter case, unrestricted global stability is proved, and in the former it is shown that a wide class of initial data (not necessarily small) generate solutions converging to stationary solutions.

A kind of nucleation procedure (outside the context of the given FBP) is described in section 9, and a discussion related to various kinds of energy is given in section 10. The paper ends with some further remarks in section 11.

2. The FBP. The PDE model consists of the following equations for functions $u = u^\epsilon(x,t),\ v = v^\epsilon(x,t),\ w = w^\epsilon(x,t)$:

$$u_t = \Delta w, \tag{1}$$

$$w = -\left(\epsilon\Delta u - \frac{1}{\epsilon}f(u) - v\right), \tag{2}$$

$$-\Delta v = u - \bar{u}. \tag{3}$$

These equations are to hold in a bounded domain Ω, and Neumann boundary conditions for v, u, and w are prescribed on $\partial\Omega$. The quantity $\bar{u}$ is the average value of u; it can be seen by integrating (1) to be independent of time. Finally, the function $f(u) = F'(u)$, where typically $F(u) = \frac{1}{2}(1-u^2)^2$, although with little change it can be taken to be any C^1 function with minima of 0 assumed only at $u = \pm 1$. The actual system studied in [5, 6, 8, 10] was like (1)–(3), but with (2) replaced by

$$w = -(\hat{\epsilon}^2\Delta u - f(u) - \sigma v), \tag{4}$$

$\sigma = O(1)$. This system can be reduced to (1)–(3) by rescaling space, time, v, and w and by defining $\epsilon = \hat{\epsilon}^{2/3}$.

As mentioned before, there is an FBP obtained in [5] as an asymptotic approximation to (1)–(3) in some sense for small ϵ. It takes the following form in two dimensions; these equations extend immediately to any other dimension.

FBP. Given a curve Γ_0 separating the planar domain Ω into two parts $\Omega_0^\pm$, find $\Gamma(t)$, $w(x,t)$, and $v(x,t)$ for $t \geq 0$, with $\Gamma(t)$ separating the plane into domains $\Omega^\pm(t)$ satisfying

$$\Delta w = 0 \quad \text{in } \Omega^\pm(t), \tag{5}$$

$$\frac{\partial w}{\partial n} = 0 \quad \text{on } \partial\Omega, \tag{6}$$

$$V = -\frac{1}{2}\left[\frac{\partial w}{\partial n}\right] \quad \text{on } \Gamma(t), \tag{7}$$

$$K = \frac{3}{2}(w - v) \quad \text{on } \Gamma(t), \tag{8}$$

$$-\Delta v = u - \bar{u} \quad \text{in } \Omega, \tag{9}$$

$$\frac{\partial v}{\partial n} = 0 \quad \text{on } \partial\Omega, \tag{10}$$

$$u = \pm 1 \text{ in } \Omega^\pm(t), \tag{11}$$

and initial condition $\Gamma(0) = \Gamma_0$, where V is the normal velocity of Γ, K is its curvature (counted positive if the center of curvature lies on the Ω^+ side), and $\bar{u}$ is the average value of u over Ω.

The directional derivative in (7) is in the direction toward Ω^+, V is considered to be positive if motion is in that same direction, and the jump indicated there is the derivative on the positive (Ω^+) side minus that on the Ω^- side.

The rigorous connection between (1)–(3) and (5)–(11) was studied by Henry [2] in the case of radial solutions in three dimensions. Among other things, beginning with a sequence of energy-bounded solutions of (1)–(3) corresponding to a sequence of values of ϵ approaching 0, Henry showed that a subsequence converges to a weak solution of the FBP. This weak solution satisfies (8) at any point where there is a spatial jump in the limit function u between -1 and 1.

3. The one-dimensional (1D) case.

In one dimension the left side of (8) vanishes, and the problem is invariant under the scaling $x \to \gamma x, t \to t, w \to \gamma^2 w,\ v \to \gamma^2 v,\ u \to u$ for arbitrary γ. This implies that any stationary spatial solution of the FBP (pattern) filling the whole line can be expanded or contracted at will, and it will still be a stationary solution with the same stability properties. This appears to contradict the selection of a preferred spacing on the basis of energy minimization, as in [9] and in [8], but does not because the FBP does not allow the creation or deletion of interfaces, so that the spacing is usually predetermined from the initial data. See sections 8 and 11 for more on this issue.

Now let Ω be the interval $(0, L)$, and let $\Gamma(t)$ consist of N distinct ordered moving points $\{x_n(t)\}$, $n = 1, 2, \ldots, N$ contained in Ω.

Let $\nu_n = 1$ if u jumps from -1 to $+1$ as x_n is traversed from left to right (so that $(x_n, x_{n+1}) \in \Omega^+$), and let $\nu_n = -1$ if the jump is from $+1$ to -1. Note that ν_n is the value of u in (x_n, x_{n+1}); since $u = \pm 1$ in alternate intervals, $-\nu_n$ is the value of u in (x_{n-1}, x_n).

Let space derivatives be denoted by " $'$ " and time derivatives by " $\dot{}$ ". We also use the notation

$$v_n = v(x_n), \quad p_n = v'(x_n).$$

The equations corresponding to (5)–(10) are

$$w'(x) = \text{ const on } (x_n, x_{n+1}) \text{ for each } n = 1, \dots, N-1, \tag{12}$$

$$w'(x) = 0 \text{ on } (0, x_1) \text{ and } (x_N, L), \tag{13}$$

$$-2\nu_n \dot{x}_n = w'(x_n + 0) - w'(x_n - 0),\ n = 1, \dots, N, \tag{14}$$

$$w(x_n) = v(x_n) \quad \text{for all } n = 1, \dots, N, \tag{15}$$

$$v'' = \bar{u} - u, \tag{16}$$

$$v'(0) = v'(L) = 0. \tag{17}$$

Note that in (14), $\dot{x}_n$ represents ν_n times the velocity V in (7), $\frac{\partial}{\partial n} = \nu_n \frac{d}{dx}$, and $\left[\frac{\partial w}{\partial n}\right] = \nu_n \left[\frac{\partial w}{\partial n}|_{x_n+0} - \frac{\partial w}{\partial n}|_{x_n-0}\right] = w'(x_n + 0) - w'(x_n - 0)$.

It will be convenient to extend this problem by reflection to the entire real line. Namely, we extend the functions $u,\ v,\ w$ to be $2L$-periodic functions which are even with respect to the points $x = 0$ and L.

In this periodic extension, each of the original points x_n (which we shall call interfacial points) has a counterpart $-x_n$ (in fact, many counterparts, by periodicity). The point $x = 0$ is not interfacial because $u = -\nu_1$ on the entire interval $(-x_1, x_1)$. A similar statement holds at $x = L$.

We must check whether the extended functions (u, v, w) continue to satisfy the above equations. It is immediate that $w' = $ const (12) in each of the new intervals between interfacial points. As for (14), we note that in going from x_n to $-x_n$, we must replace ν_n by $-\nu_n$, V by $-V$, and $\frac{\partial}{\partial n}$ by $-\frac{\partial}{\partial n}$. Making these replacements changes (14) into

$$-2(-\nu_n)(-\dot{x}_n) = -[w'(-x_n - 0) - w'(-x_n + 0)],$$

which is still seen to be the form that (7) takes at the interfacial point $-x_n$. Therefore, (7) continues to be satisfied by the extended functions. Finally, (15) and (16) are still valid.

Thus each solution of the problem on $(0, L)$ gives rise to a $2L$-periodic solution on the whole line, even with respect to 0 and L. Conversely, any such solution on the whole line, when restricted to $(0, L)$, is a solution of the finite interval problem; in fact, the evenness implies the boundary conditions (13) and (17). Thus the two problems are equivalent.

In view of this, we shall usually confine our attention in the following to problems on the whole line. In fact we no longer necessarily assume periodicity. The interfacial points $\{x_n\}$ will have indices n ranging over all positive and negative integers: $-\infty < n < \infty$.

Integrating (16), we have, for all n,

$$v'(x) = p_n + (\bar{u} - \nu_n)(x - x_n), \quad x \in [x_n, x_{n+1}], \tag{18}$$

$$v(x) = v_n + p_n(x - x_n) + \frac{1}{2}(\bar{u} - \nu_n)(x - x_n)^2. \tag{19}$$

We now denote the interval lengths by $\xi_n = x_{n+1} - x_n$. From (18) and (19),

$$\frac{v(x_{n+1}) - v(x_n)}{\xi_n} = p_n + \frac{1}{2}(\bar{u} - \nu_n)\xi_n, \tag{20}$$

$$p_{n+1} - p_n = (\bar{u} - \nu_n)\xi_n. \tag{21}$$

And from (12) and (15),

$$w'(x) = \frac{v(x_{n+1}) - v(x_n)}{\xi_n}, \quad x \in (x_n, x_{n+1}). \tag{22}$$

Hence from (22), (20), and (14),

$$-2\nu_n \dot{x}_n = p_n - p_{n-1} + \frac{1}{2}\left[(\bar{u} - \nu_n)\xi_n - (\bar{u} - \nu_{n-1})\xi_{n-1}\right]. \tag{23}$$

Substituting (21) into (23) and multiplying by $\nu_n/2$, we find

$$-\dot{x}_n = \frac{1}{4}[(\nu_n\bar{u} - 1)\xi_n + (\nu_n\bar{u} + 1)\xi_{n-1}], \tag{24}$$

and hence

$$\dot{\xi}_n = \frac{1}{4}\left[(\nu_n\bar{u} + 1)\xi_{n+1} + 2(\nu_n\bar{u} - 1)\xi_n + (\nu_n\bar{u} + 1)\xi_{n-1}\right]. \tag{25}$$

Our main effort will be devoted to solving the initial-value problem associated with (25). After the solution is found, it is straightforward to obtain $x_0(t)$ by solving (from (24))

$$\dot{x}_0 = -\frac{1}{4}\left[(\nu_0\bar{u} - 1)\xi_0 + (\nu_0\bar{u} + 1)\xi_{-1}\right], \tag{26}$$

and hence

$$x_n(t) = x_0(t) + \sum_0^{n-1} \xi_j(t) \tag{27}$$

for all $n > 0$, a similar representation holding for $n < 0$.

The evolution (25) is especially revealing in the case when $\bar{u} = 0$, for then it is a space-discretization of the ordinary heat equation

$$\dot{\xi}_n = \frac{1}{4}\left[\xi_{n+1} - 2\xi_n + \xi_{n-1}\right], \tag{28}$$

in which the index n plays the role of space, and ξ_n plays the role of temperature at location n.

In the general case when $\bar{u} \neq 0$, the coefficients on the right side of (25) depend on n, which in a sense represents the discretization of a heat equation for an inhomogeneous medium, the diffusivity being a function of the grid point which alternates between one positive value and another.

If one thinks of the heat equation $U_t = U_{xx}$ as being an efficient way to evolve an initial function $U(x, 0)$ of x toward a constant function equal to the spatial average of $U(x, 0)$ (and it is, in the sense of being a steepest-descent rule for decreasing $\int U_x(x, t)^2 dx$), then (28) is the analogous efficient way to evolve a sequence of interfacial points $\{x_n(0)\}$ to a configuration in which they are equally spaced, so that the ξ_n are independent of n. A similar interpretation can be given to the case when $\bar{u} \neq 0$. The outcome of this evolution process is studied carefully in sections 7 and 8.

Returning to problems on a finite interval $[0, L]$ with N interfaces, we note that the even extension to the whole line described above generates a solution of (25), defined for all n, such that ξ_n is periodic in n of period $2N$. In fact each interval length $\xi_n = x_{n+1} - x_n$ for $n = 1, 2, \dots, N - 1$ has by reflection its counterpart $(-x_n) - (-x_{n+1})$ (the length of a subinterval of $[-L, 0]$), which we may rename ξ_{-n}. In addition, there is the interval $-x_1, x_1$ (whose length we rename ξ_0) spanning the origin, and the interval length ξ_N spanning the point $x = L$. This makes a total of $2N$ subintervals covering a basic period interval $(-x_{N-1}, 2L - x_{N-1})$ of length $2L$. The periodic extension simply involves translating the basic period interval, with its $2N$ subintervals, periodically to cover the whole line.

In section 4.7 we prove the existence, given initial data $\xi_n(0)$ which are $2N$-periodic in n, of a solution which remains $2N$-periodic as time evolves. This, together with the uniqueness of solutions of the general initial-value problem, implies that periodic initial data always generate periodic solutions.

4. General solution. The purpose of this section is to provide an explicit solution of the initial-value problem for (25).

The number $\bar{u}$ in (25) was originally introduced in (3) as the spatial average of u in a bounded domain. However, the system (25) has meaning independently of the significance of $\bar{u}$, and we consider its solution for arbitrary $\bar{u} \in (-1, 1)$. It will be seen in (68) that stationary solutions are such that $\bar{u}$ is indeed the average of u, and for solutions on a finite interval, we see in section 8 that the evolution leads to a final state with this same property.

For definiteness, we take

$$\nu_n = (-1)^n$$

throughout the rest of the paper, so that $u = 1$ on (x_n, x_{n+1}) for n even.

4.1. Formal solution in the case $\bar{u} = 0$. This case is straightforward and involves solving (28) by Fourier transform. We first derive the solution in a formal manner; the justification, along with that of section 4.2, will follow in section 4.5.

We represent

$$\xi_n(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \tilde{\alpha}(k, t) e^{ink} dk, \tag{29}$$

the inverse transform being

$$\tilde{\alpha}(k,t) = \sum_{n=-\infty}^{\infty} \xi_n(t) e^{-ikn}. \tag{30}$$

The system (28) becomes

$$\frac{\partial \tilde{\alpha}}{\partial t} = \frac{1}{2}(\cos k - 1)\tilde{\alpha}, \tag{31}$$

whose solution is

$$\tilde{\alpha}(k,t) = \tilde{A}(k) \exp\left[-\frac{1}{2}(1-\cos k)t\right], \tag{32}$$

and the spectral function $\tilde{A}(k)$ can be obtained from the initial data $\{\xi_n(0)\}$. We therefore have

$$\xi_n(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \tilde{A}(k) \exp\left[-\frac{1}{2}(1-\cos k)t\right] e^{ink} dk. \tag{33}$$

Conditions under which these equations, and those given below, have meaning in the distribution sense will be given in section 4.5.

4.2. Formal solution in the case $\bar{u} \neq 0$. We split the numbers ξ_n into two sets, corresponding to even and odd n, and then rescale, so defining

$$\zeta_n = (1-\bar{u})\xi_{2n}, \qquad \eta_n = (1+\bar{u})\xi_{2n+1}. \tag{34}$$

From (25) we have

$$\frac{1}{1-\bar{u}}\dot{\zeta}_n = \dot{\xi}_{2n} = \frac{1}{4}\left[(\bar{u}+1)\xi_{2n+1} - 2(1-\bar{u})\xi_{2n} + (1+\bar{u})\xi_{2n-1}\right]$$

$$= \frac{1}{4}(\eta_n - 2\zeta_n + \eta_{n-1}). \tag{35}$$

Similarly,

$$\frac{1}{1+\bar{u}}\dot{\eta}_n = \frac{1}{4}(\zeta_{n+1} - 2\eta_n + \zeta_n). \tag{36}$$

Again, the system (35), (36) may be solved by Fourier transform. We represent

$$\zeta_n(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \alpha(k,t) e^{ink} dk, \quad \eta_n(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \beta(k,t) e^{ink} dk, \tag{37}$$

the inverse transforms being

$$\alpha(k,t) = \sum_{n=-\infty}^{\infty} \zeta_n(t) e^{-ikn}, \quad \beta(k,t) = \sum_{n=-\infty}^{\infty} \eta_n(t) e^{-ikn}. \tag{38}$$

The system (35), (36) becomes

$$\frac{1}{1-\bar{u}}\dot{\alpha} = \frac{1}{4}\left[\beta(1+e^{-ik}) - 2\alpha\right],$$

$$\frac{1}{1+\bar{u}}\dot{\beta} = \frac{1}{4}\left[\alpha(1+e^{ik}) - 2\beta\right]. \tag{39}$$

The general solution of this linear system is

$$\begin{aligned}\alpha(k,t) &= A(k)e^{\sigma_+(k)t} + \lambda_-(k)B(k)e^{\sigma_-(k)t},\\ \beta(k,t) &= \lambda_+(k)A(k)e^{\sigma_+(k)t} + B(k)e^{\sigma_-(k)t}\end{aligned} \tag{40}$$

for arbitrary $A,\ B$, which will be allowed to be distributions (see section 4.4 below), where

$$\sigma_\pm(k) = \frac{1}{2}\left[-1 \pm \sqrt{\bar{u}^2 + (1-\bar{u}^2)\cos^2(k/2)}\right], \tag{41}$$

$$\lambda_\pm(k) = \frac{2(2\sigma_\pm + 1 \mp \bar{u})}{(1\mp\bar{u})(1+e^{\mp ik})}. \tag{42}$$

It is seen, since $\bar{u} \neq 0$, that the $\sigma_\pm$ and $\lambda_\pm$ are real, C^∞ is periodic in k with period 2π, and for $-\pi < k \leq \pi$,

$$\sigma_-(k) \leq -\frac{1}{2},\ \sigma_+(k) < 0 \text{ when } k \neq 0, \quad \sigma_+(0) = 0. \tag{43}$$

For small k, we have

$$\sigma_+(k) \sim -\frac{1}{16}(1-\bar{u}^2)k^2. \tag{44}$$

Note that the numerator and denominator in (42) both vanish when $k \to \pm\pi$, but the ratio remains bounded.

In all, we have

$$\zeta_n(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[A(k)e^{\sigma_+(k)t} + \lambda_-(k)B(k)e^{\sigma_-(k)t}\right]e^{ink}dk, \tag{45}$$

$$\eta_n(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\lambda_+(k)A(k)e^{\sigma_+(k)t} + B(k)e^{\sigma_-(k)t}\right]e^{ink}dk. \tag{46}$$

The distributions A and B can be found from initial values $\xi_n(0)$ by setting $t=0$ in (40), (37) and inverting. Specifically,

$$A(k) = (1-\lambda_-(k)\lambda_+(k))^{-1}\sum_n\left[\zeta_n(0) - \lambda_-(k)\eta_n(0)\right]e^{-ikn}, \tag{47}$$

$$B(k) = (1-\lambda_-(k)\lambda_+(k))^{-1}\sum_n\left[\eta_n(0) - \lambda_+(k)\zeta_n(0)\right]e^{-ikn}. \tag{48}$$

4.3. Real solutions. If we should wish to consider all complex-valued solutions of (35) and (36), $A(k)$ and $B(k)$ are arbitrary. However, our motivation dictates that the solution be real and positive. Of course, the real and imaginary parts of any complex solution of (35), (36) are also solutions, so we shall be justified in studying complex ones. Apart from that, it may be of interest to characterize those distributions A and B which generate real solutions.

We shall show that ζ_n and η_n are real for all n if and only if

$$\overline{A(-k)} = A(k), \quad \overline{B(-k)} = B(k) \tag{49}$$

for all k. Taking the complex conjugate of (45), we obtain

$$\overline{\zeta_n(t)} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\overline{A(k)}e^{-ink}e^{\sigma_+(k)t} + \overline{\lambda_-(k)B(k)}e^{-ink}e^{\sigma_-(k)t}\right]dk. \tag{50}$$

By transforming the integration variable $k\rightarrow -k$, noting that $\sigma_\pm(-k) = \sigma_\pm(k)$ and $\lambda_\pm(-k) = \overline{\lambda_\pm(k)}$, we obtain

$$\overline{\zeta_n(t)} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\overline{A(-k)}e^{ink}e^{\sigma_+(k)t} + \lambda_-(k)\overline{B(-k)}e^{ink}e^{\sigma_-(k)t}\right]dk.$$

The condition that ζ_n must be real is that $\overline{\zeta_n(t)} = \zeta_n(t)$. Comparing this last representation to (45), we see that the condition reduces to (49). The same analysis can be applied to show that the reality of the η_n is also equivalent to (49).

By the use of (49), we have the representation

$$\zeta_n(t) = \frac{1}{\pi}\int_0^{\pi} \operatorname{Re}\left[e^{ink}\left(A(k)e^{\sigma_+(k)t} + \lambda_-(k)B(k)e^{\sigma_-(k)t}\right)\right]dk, \tag{51}$$

$$\eta_n(t) = \frac{1}{\pi}\int_0^{\pi} \operatorname{Im}\left[e^{ink}\left(\lambda_+(k)A(k)e^{\sigma_+(k)t} + B(k)e^{\sigma_-(k)t}\right)\right]dk. \tag{52}$$

This time the distributions A and B, defined only for $k \in [0,\pi]$, need only satisfy (49) when $k = 0$; otherwise, they are arbitrary.

4.4. Periodic distributions. We shall show (Theorem 1) that if the initial values $\xi_n(0)$ satisfy a growth condition with respect to n, then the above formal analysis is justified, i.e., a global solution of the initial-value problem exists in the form (45)–(48).

But first, we briefly review the Fourier transform theory for distributions.

We define a 2π-periodic distribution α to be a linear functional on the space of C^∞ 2π-periodic functions (test functions) $\phi(k)$, its value being denoted by $\langle\alpha,\phi\rangle$, which is continuous in the sense that if $\{\phi_j\}$ is a sequence of test functions such that for each $m = 0, 1, \ldots$

$$\lim_{j\to\infty}\left(\frac{d}{dk}\right)^m \phi_j(k) = \left(\frac{d}{dk}\right)^m \phi(k)$$

uniformly in k, then

$$\lim_{j\to\infty}\langle\alpha,\phi_j\rangle = \langle\alpha,\phi\rangle.$$

When $\alpha(k)$ is an integrable function, we use the convention that $\langle\alpha,\phi\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi}\alpha(k)$ $\phi(k)\,dk$.

Thus the right side of (45) is to be interpreted as

$$\langle A, \exp[\sigma_+(k)t + ink]\rangle + \langle B, \lambda_-(k)\exp[\sigma_-(k)t + ink]\rangle$$

if A and B are distributions. The right side of (46) is interpreted similarly.

The continuity of distributions implies that if $\phi(k,t)$ is a test function for each t such that for all m, $\frac{\partial^m \phi}{\partial k^m}$ is continuously differentiable in t, then

$$\frac{d}{dt}\langle A, \phi(\cdot,t)\rangle = \left\langle A, \frac{\partial \phi}{\partial t}(\cdot,t)\right\rangle. \tag{53}$$

In fact, we just express the t-derivatives as limits of difference quotients.

It is this property (53) that ensures, for any distributions A and B, that the functions $\zeta_n(t)$, $\eta_n(t)$ given by (45), (46) satisfy (35), (36). We just differentiate (45), (46) and use (39), (40).

4.5. Justification.

THEOREM 1. *Assume that for some* $p > 0$, $C > 0$ *the initial data satisfy* $|\xi_n(0)| \le C(1+|n|^p)$ *for all* n. *Then* (47), (48) *define distributions A, B, and* (45), (46) *provide the solution of the initial-value problem for* (35), (36) *for all* $t \ge 0$.

Proof. We have just seen that if A and B are any distributions, then (45), (46) provide a solution of (35), (36). It remains only to show (i) that the distributions A and B can be obtained in terms of the initial conditions by the expressions (47), (48) provided the right sides of those equations really are distributions, and (ii) that the right sides are distributions provided that the initial data satisfy the growth condition. The first task (i) is simply a consequence of the inversion formula. Therefore, we recall the proof of that formula, which takes this form: If we set $e_n(k) = e^{ink}$ and, for any given distribution α, set $\alpha_n = \langle \alpha, e_n\rangle$, then α can be reconstituted as

$$\alpha = \sum_n \alpha_n \overline{e_n}. \tag{54}$$

The meaning of (54) is simply that for every test function ϕ,

$$\langle \alpha, \phi\rangle = \sum_n \alpha_n \langle \overline{e_n}, \phi\rangle. \tag{55}$$

The proof of this follows from the standard inversion formula for smooth functions: Let ϕ be any test function. If we set $\phi_n = \langle \overline{e_n}, \phi\rangle$, then

$$\phi(k) = \sum_n \phi_n e_n(k), \tag{56}$$

and the convergence is uniform. Applying the distribution α to (56) with use of the continuity property of α, we get precisely (55), which completes the proof of (i).

Finally, we show (ii) that the stated growth condition implies that the right sides of (47) and (48) are distributions. We do this for (47), as the argument for (48) is the same. First, the growth condition on the $\xi_n(0)$ implies

$$|\zeta_n(0)| + |\eta_n(0)| \le C(1+|n|^{2p}). \tag{57}$$

(Here C is a generic constant.) Now let ϕ be any test function, and let R_n be the nth term on the right of (47). We have the estimate

$$|\langle R_n, \phi\rangle| \le C(1+|n|^{2p})|\langle \phi, \overline{e_n}\rangle|. \tag{58}$$

But $|\langle \overline{e_n}, \phi\rangle|$ decays as $|n|\to\infty$ faster than any negative power of $|n|$, by virtue of the infinite differentiability of ϕ (integrate by parts any number of times). Therefore, using (58), we see that the series $[1-\lambda_-(k)\lambda_+(k)]^{-1}\sum[\zeta_n(0)-\lambda_-(k)\eta_n(0)]\langle\overline{e_n},\phi\rangle$ corresponding to the right side of (47) converges uniformly in k. We call the limit $\langle A,\phi\rangle$. We omit the proof that this linear functional is continuous in the right sense, so that it defines a distribution A. □

4.6. Monochromatic solutions. By way of example, consider the case when the initial data are

$$\xi_n(0) = e^{i\gamma n} \tag{59}$$

for some real number γ, which (since $e^{i\gamma n}$ is 2π-periodic in γ) we may and shall confine to the interval $0 \le \gamma < 2\pi$. Then

$$\zeta_n(0) = (1-\bar{u})e^{i2\gamma n}, \quad \eta_n(0) = (1+\bar{u})e^{i\gamma}e^{i2\gamma n}. \tag{60}$$

We can then find A and B from (47), (48) and the solution from (45), (46). However, we take a more direct approach, supposing that the solution, in its dependence on n, remains monochromatic, proportional to $e^{i2\gamma n}$. Thus in the spectral representation (45), (46), only the value $k = 2\gamma$ appears. (It of course may happen that 2γ lies outside the range of integration $[-\pi,\pi]$ in (45), (46). However, $\sigma_\pm$, A, B, and $\lambda_\pm$ are all 2π-periodic functions of k, so we really mean the value of k (mod 2π) which does lie in $[-\pi,\pi]$.) The distributions A and B will be delta-functions.

Prompted by these considerations, we propose to find a solution in the form

$$\zeta_n(t) = a_+(\gamma)\exp\left(\sigma_+(2\gamma)t + 2in\gamma\right) + \lambda_-(2\gamma)a_-(\gamma)\exp\left(\sigma_-(2\gamma)t + 2in\gamma\right), \tag{61}$$

$$\eta_n(t) = \lambda_+(2\gamma)a_+(\gamma)\exp\left(\sigma_+(2\gamma)t + 2in\gamma\right) + a_-(\gamma)\exp\left(\sigma_-(2\gamma)t + 2in\gamma\right) \tag{62}$$

for some constants $a_\pm(\gamma)$. We set $t = 0$ and replace the left sides of (61), (62) by the right sides of (60). The resulting equations can be solved for a_+ and a_-:

$$\begin{aligned} a_+(\gamma) &= [1-\lambda_+(2\gamma)\lambda_-(2\gamma)]^{-1}[(1-\bar{u}) - \lambda_-(2\gamma)(1+\bar{u})e^{i\gamma}], \\ a_-(\gamma) &= [1-\lambda_+(2\gamma)\lambda_-(2\gamma)]^{-1}[(1+\bar{u}))e^{i\gamma} - \lambda_+(2\gamma)(1-\bar{u})]. \end{aligned} \tag{63}$$

It is now straightforward to verify that (61), (62) is indeed the desired solution.

4.7. The solution for a finite interval. It was shown in section 3 that the problem with N interfacial points on a finite interval is equivalent to a problem on the whole real line, with the solution ξ_n required to be periodic in n of period $2N$.

First of all, we represent the $2N$-periodic initial data $\xi_n(0)$ in the form

$$\xi_n(0) = \sum_{m=0}^{2N-1} c_m e^{i\frac{2\pi}{2N}mn}. \tag{64}$$

These are $2N$ equations ($n = 0,\ldots,2N-1$) for the $2N$ unknowns c_m. They can be solved by forming the combination $\sum_{n=0}^{2N-1}\xi_n(0)e^{-i\frac{2\pi}{2N}nm}$ and observing the relation

$$\sum_{n=0}^{2N-1} e^{i\frac{2\pi}{2N}np} = \begin{cases} 2N, & p = 0, \\ 0, & p = 1,2,\ldots,2N-1. \end{cases} \tag{65}$$

We find that (64) holds with

$$c_m = \frac{1}{2N} \sum_{n=0}^{2N-1} \xi_n(0) e^{-i\frac{2\pi}{2N}nm}. \tag{66}$$

From (64) and the linearity property, we conclude that the solution $\xi_n(t)$ is a linear combination of monochromatic solutions as in (61), (62), (63) with $\gamma = \frac{2\pi m}{2N} = \frac{\pi m}{N} \equiv \gamma_m$, $m = 0, \ldots, 2N-1$, and coefficients c_m (66).

5. Positivity of the $\boldsymbol{\xi_n}$. Suppose $\xi_n(0) > 0$ for all n. We show that it is impossible for a finite number of these interval lengths to vanish at some finite positive $t = T > 0$, the surrounding ones remaining positive. Specifically, given some finite nonempty sequence $S = \{n : n_1 \le n \le n_2\}$ (the simplest case is when $n_1 = n_2$), we say that this set collapses at time $T > 0$ if (i) $\lim_{t\uparrow T} \xi_n(t) = 0$ and $\xi_n(t) > 0$ for $t < T$ close to T for $n \in S$, and (ii) $\xi_n(t) > 0$ for $t \le T$ close to T for $n = n_1 - 1$ or $n_2 + 1$.

THEOREM 2. *If* $\xi_n(t)$ *satisfy* (25)*, no finite sequence collapses at any finite time.*

Proof. If S collapses at time T, then for $t < T$ close to T, ξ_{n_2} is arbitrarily small, but ξ_{n_2+1} is bounded away from zero. Therefore, it follows from (25) with $n = n_2$ that $\dot{\xi}_{n_2} > 0$, contradicting its approach to 0. □

This leads to the following global existence result.

COROLLARY 1. *Any positive initial data satisfying the hypotheses of Theorem* 1 *give rise to a global positive solution of* (25) or (35), (36).

6. Stationary and similarity solutions. First, we exhibit the stationary solutions on the whole line. From (35), (36) with the left side set equal to 0, we derive the result that $\zeta_{n+1} - \zeta_n \equiv \alpha$ and $\eta_{n+1} - \eta_n \equiv \beta$ are independent of n. Therefore, $\zeta_n = \zeta_0 + n\alpha$, $\eta_n = \eta_0 + n\beta$. But, again using (35), we find $\alpha = \beta$ and $\eta_0 = \zeta_0 - \frac{\alpha}{2}$, so that

$$\zeta_n = \zeta_0 + n\alpha, \quad \eta_n = \zeta_0 + \left(n - \frac{1}{2}\right)\alpha. \tag{67}$$

This provides a two-parameter family of solutions, the parameters being α and ζ_0. But since we want solutions which are positive for each n, we should select $\alpha = 0$.

In the case when $\alpha = 0$, we obtain constant solutions

$$(1 - \bar{u})\xi_{2n} = \zeta_n = \eta_n = (1 + \bar{u})\xi_{2n+1} \equiv \zeta_0 \quad \text{for all } n.$$

For these constant solutions, the interval lengths alternate between $\frac{\zeta_0}{1-\bar{u}}$ and $\frac{\zeta_0}{1+\bar{u}}$. The "average" value of u can then be verified to be

$$\frac{\xi_{2n} - \xi_{2n+1}}{\xi_{2n} + \xi_{2n+1}} = \frac{\frac{1}{1-\bar{u}} - \frac{1}{1+\bar{u}}}{\frac{1}{1-\bar{u}} + \frac{1}{1+\bar{u}}} = \bar{u}. \tag{68}$$

It may be checked that these stationary solutions correspond to the choice

$$A(k) = 2\pi\zeta_0\delta(k), \quad B(k) = 0 \tag{69}$$

in (45), (46). The stationary solutions shown above, which grow linearly in n and which we had to reject, correspond to $A(k)$ being an imaginary constant times $\delta'(k)$ and $B(k) = 0$. Positive solutions which grow quadratically in n may also be constructed; they are not stationary, but rather they grow linearly in t. For them, the function $A(k)$ is a constant times $\delta''(k)$. This process may be continued to obtain similarity solutions corresponding to any derivative of $\delta(k)$.

7. Stability for problems on the whole line. Given a stationary solution $\{\xi_n^0\}$ (or any other solution, for that matter), we now endeavor to characterize initial data $\xi_n(0)$ such that the resulting evolution $\xi_n(t)$ will converge to $\{\xi_n^0\}$ as $t\to\infty$. After that, we consider conditions under which the corresponding result holds for the interface locations $x_n(t)$.

Since our basic problem (25) is linear and homogeneous, it suffices to find solutions which decay to 0 as $t\to\infty$; those solutions (which could now be negative) will then be considered as perturbations of the basic solution $\{\xi_n^0\}$, and the only additional restriction we might wish to impose is that the basic solution plus the perturbation be real and positive initially (hence at all later times).

We give two characterizations of decaying solutions.

THEOREM 3. *Let* $\xi_n(0)\to 0$ *as* $|n|\to\infty$. *Then*

$$\max_n |\xi_n(t)|\to 0 \text{ as } t\to\infty. \tag{70}$$

Remark. Part of the proof is a simple comparison argument for the system (35), (36), which is quasi-monotone. More detailed results based on monotonicity arguments can be derived.

THEOREM 4. *Assume the spectral pair* $A(k)$, $B(k)$, *obtained from the initial data by* (47), (48), *satisfies any one of the following, or is a linear combination of pairs, each satisfying one of the following. Then* $\lim_{t\to\infty}\xi_n(t) = 0$ *uniformly in* n. *In case* 1, *the convergence is exponential; in the second case, it may be algebraic of order* $t^{(\alpha-1)/2}$.

1. $A(k) = c_1\delta(k-k_0)$ *and* $B(k) = c_2\delta(k-k_1)$ *for some constants* c_i *and* k_i *with* $k_0 \neq 0$.
2. $|A(k)| + |B(k)| \le c|k|^{-\alpha}$ *for some constants* c *and* $\alpha < 1$.

Remark. In case 2, Theorem 3 applies, because it can be shown that the initial values decay as $|n|\to\infty$. Theorem 4 gives the rate of decay.

We prove Theorem 4 first and then Theorem 3.

Proof of Theorem 4. In the first case, we get from (45) that $\zeta_n(t) = \frac{c_1}{2\pi}e^{\sigma_+(k_0)t}e^{ink_0} + \frac{c_2\lambda_-(k_1)}{2\pi}e^{\sigma_-(k_1)t}e^{ink_1}$, and the conclusion follows for the ζ_n because, by hypothesis, $\sigma_+(k_0)$ and $\sigma_-(k_1)$ are negative. The same argument works for the η_n as well.

This result shows that for time decay it is not necessary for the initial values of the ξ_n to approach 0 as $|n|\to\infty$; examples are the monochromatic solutions in section 4.6 with $\gamma\neq 0$.

In the second case, we again use (45), and (44) as well, to obtain for some $c, \beta > 0$

$$|\zeta_n(t)| < c\int_{-\infty}^{\infty} |k|^{-\alpha}e^{-\beta k^2 t}dk$$

$$= [\text{setting } \kappa = k\sqrt{t}] \quad c\int_{-\infty}^{\infty} t^{(\alpha-1)/2}|\kappa|^{-\alpha}e^{-\beta\kappa^2}d\kappa = ct^{(\alpha-1)/2}, \tag{71}$$

and the conclusion again follows.

LEMMA 1. *Let* $\xi_n(0)\to 0$ *as* $|n|\to\infty$. *Then for all* $t \ge 0$, $X(t) \equiv \max_n \max[|\zeta_n(t)|, |\eta_n(t)|] \le X(0)$.

Proof. Defining

$$z_n = \begin{cases} (1-\bar{u})\xi_n, & n \text{ even}, \\ (1+\bar{u})\xi_n, & n \text{ odd}, \end{cases}$$

we represent $X(t) = \max_n |z_n(t)|$ and write (35), (36) (or (25)) in the form

$$\frac{4}{1-(-1)^n\bar{u}}\dot{z}_n = z_{n+1} - 2z_n + z_{n-1}. \tag{72}$$

For any $\epsilon > 0$, let $y_n = z_n - \epsilon n^2 - 2\epsilon t$, so that from (72)

$$\frac{4}{1-(-1)^n\bar{u}}\dot{y}_n = y_{n+1} - 2y_n + y_{n-1} + 2\epsilon - \frac{8\epsilon}{1-(-1)^n\bar{u}}. \tag{73}$$

We have $y_n(0) \le X(0)$ for all n and (since the z_n are bounded) $y_n(t) < 0$ for large n. If $y_n(t^*) > X(0)$ for some $n,\ t^* > 0$, $\max_{n,t\in[0,t^*]} y_n(t)$ is achieved at some $n = n_1,\ t = t_1$, where $\dot{y}_{n_1}(t_1) \ge 0,\ \ y_{n_1+1}(t_1) - 2y_{n_1}(t_1) + y_{n_1-1}(t_1) \le 0$. Substituting these inequalities into (73), we obtain that the left side is ≥ 0 and the right side is $\le -2\epsilon < 0$. This contradiction implies that $y_n(t) \le X(0)$ for all $n,\ t$.

Since this is true for every choice of $\epsilon > 0$, we conclude that $z_n(t) \le X(0)$. Similarly, we get that $z_n(t) \ge -X(0)$, so that $X(t) \le X(0)$. □

LEMMA 2. *If $\xi_n(0) = 0$ for $|n|$ large enough, then $\xi_n(t)\to 0$ as $t\to\infty$, uniformly in n.*

Proof. In (47) and (48) there are only a finite number of terms in the summations on the right. Therefore, $A(k)$ and $B(k)$ are bounded functions of k, and the second hypothesis of Theorem 4 holds with $\alpha = 0$. □

Proof of Theorem 3. Given any $\epsilon > 0$, let $N(\epsilon)$ be such that $|\xi_n(0)| < \epsilon$ for $|n| > N(\epsilon)$. Write $\xi_n(t) = \xi_n^1(t) + \xi_n^2(t)$, where $\xi_n^1(t)$ is the evolution starting from

$$\xi_n^1(0) = \begin{cases} \xi_n(0), & |n| \le N(\epsilon), \\ 0, & |n| > N(\epsilon). \end{cases}$$

By Lemma 2, $\xi_n^1(t)\to 0$ as $t\to\infty$, so let $T(\epsilon)$ be such that $|\xi_n^1(t)| < \epsilon$ for $t > T(\epsilon)$, for all n. Since $|\xi_n^2(0)| < \epsilon$, we have from Lemma 1 that $|\xi_n^2(t)| < \epsilon$ for all n and t. Therefore, $|\xi_n(t)| < 2\epsilon$ for $t > T(\epsilon)$ and all n. This establishes the convergence.

THEOREM 5. *Assume condition* 2 *of Theorem* 4 *holds with $\alpha < 0$. Then for each m, the limit $\lim_{t\to\infty} x_m(t) = x_m(\infty)$ exists. The limit is a stationary configuration.*

Proof. We may use (26), (34), (37), (40), (44) to obtain an integral representation of $\dot{x}_0(t)$ in terms of A and B. For the convergence of $x_0(t)$ to a limit, it suffices that $\int_1^\infty |\dot{x}_0(t)|\,dt < \infty$.

Proceeding as in the proof of Theorem 4 leads to the estimate $|\dot{x}_0(t)| \le ct^{\frac{\alpha}{2}-1}$. This is integrable as required if $\alpha < 0$. □

8. Global stability in the case of a finite interval. We showed in section 4.7 that in this case, the evolution reduces to that of a finite linear combination of monochromatic evolutions. So we briefly return to the question of the decay of monochromatic solutions considered in section 4.6. We give the argument only for the more difficult case $\bar{u} \ne 0$.

In view of (43), it is seen from (61), (62) that every monochromatic solution decays exponentially to zero unless $\sigma_+(2\gamma) = 0$, i.e., $\gamma = 0$ or π. The rate is $e^{\sigma_+(2\gamma)t}$. We now consider these two exceptional cases in more detail, beginning with the case $\gamma = 0$. We calculate from (41), (42)

$$\sigma_+(0) = 0,\quad \sigma_-(0) = -1,\quad \lambda_+(0) = 1,\quad \lambda_-(0) = -\frac{1-\bar{u}}{1+\bar{u}}. \tag{74}$$

Thus from (63)

$$a_+ = 1 - \bar{u}^2, \quad a_- = \bar{u}(1+\bar{u}), \tag{75}$$

and by (61), (62)

$$\begin{aligned} \zeta_n(t) &= (1-\bar{u}^2) - \tfrac{1-\bar{u}}{1+\bar{u}} e^{-t}, \\ \eta_n(t) &= (1-\bar{u}^2) + \bar{u}(1+\bar{u})e^{-t}. \end{aligned} \tag{76}$$

In the limit as $t \to \infty$, therefore, ζ_n and η_n approach the limit $1-\bar{u}^2$ exponentially, independently of n; hence

$$\xi_{2n}(\infty) = 1+\bar{u}, \quad \xi_{2n+1}(\infty) = 1-\bar{u}, \tag{77}$$

so that the limiting sizes of the intervals alternate between $1+\bar{u}$ and $1-\bar{u}$. It is easily verified (section 6) that the average value of u in this configuration is $\bar{u}$. (Note again that we have not assumed that the initial data for this solution have this value for their average; the number $\bar{u}$ is, for the purpose of the above calculations, just a parameter in the differential equations.)

Finally, consider the last case, $\gamma = \pi$. We have $\sigma_+(2\pi) = 0$. By periodicity, the values of the σ's and λ's are the same as in (74), but the factor $e^{i\gamma}$ in (63) is now -1. It follows that $a_+ = 0$ and $a_- = -(1+\bar{u})$. Because $a_+ = 0$, this monochromatic solution decays to 0 exponentially.

Therefore, the only one which does not is the one with $\gamma = 0$.

Therefore, in the linear combination of monochromatic solutions giving the solution of the problem on a finite interval, arising from the representation (64) of the initial data, all terms decay to zero except the one with $m = 0$. That case was analyzed in (76) above. We conclude that if (64) holds, then

$$\xi_{2n}(\infty) = c_0(1+\bar{u}), \quad \xi_{2n+1}(\infty) = c_0(1-\bar{u}). \tag{78}$$

Finally, by (66), $c_0 = \frac{1}{2N}\sum_{n=0}^{2N-1} \xi_n(0) = \frac{L}{N} \equiv \bar{\xi}$, where L is the length of the finite interval we are considering and $\bar{\xi}$ is the initial (and also final) average of the ξ_n.

In short, we have found that, irrespective of the initial values of the N intervals making up the basic x-interval of length L, the lengths of the intervals evolve to become semiuniform, i.e., the even ones all have the same length $\bar{\xi}(1+\bar{u})$, and the odd ones as well, with lengths $\bar{\xi}(1-\bar{u})$. Those two lengths are such that the average value of u is $\bar{u}$, and the total length is L. This final stationary solution is therefore globally stable.

Theorem 6. *Given any finite N and any initial data* (64), *the solution of the initial-value problem on a finite interval satisfies*

$$\xi_{2n}(\infty) = \bar{\xi}(1+\bar{u}), \quad \xi_{2n+1}(\infty) = \bar{\xi}(1-\bar{u}).$$

In view of the exponential convergence of the ξ's and the proof of Theorem 5, we also have convergence of the interfaces $x_n(t)$, although not necessarily to the locations of the original unperturbed locations; generally there will be a shift.

9. Nucleation. Nucleation in this context is the opposite of collapsing (see section 5). It involves inserting two (or more) new spurious interfaces enclosing a spurious interval of 0 length which then grows. Given a solution $\{x_n(t)\}$, an arbitrary time $t^* > 0$, and an arbitrary point x^* in the interior of one of the intervals (say, (x_0, x_1)

for definiteness) at time t^*, one may consider a new initial-value problem for interval lengths $\xi_n^*(t)$ starting at t^* with initial values of $\xi_n^*(t^*)$ obtained as follows. Set $\xi_0^*(t^*) = 0$; this corresponds to beginning with an interval of 0 length located at the point x^*. Also set $\xi_1^*(t^*) = x_1 - x^*$ and $\xi_{-1}^*(t^*) = x^* - x_0$, and require the remaining $\xi_n^*(t^*)$ to match with some $\xi_m(t^*)$ for the appropriate m, depending on n.

An argument similar to that in the proof of Theorem 2 shows that at the initial time, the derivative $\dot{\xi}_0^*(t^*) > 0$, so that the "ghost interval" ξ_0^* immediately develops positive length, making it a real interval. The new initial-value problem then has a global positive solution.

In an easy generalization of this procedure, one can think of the point x^* as representing a collection of any positive odd number of ghost intervals; then again, they all immediately become intervals of positive length. Finally, any interface located at position x^* at time t^* can be considered to be an interface abutting an even number of ghost interfaces on one side, which again move apart to generate an even number of new intervals.

10. Energy considerations. We consider the FBP on a finite interval $(0, L)$ with N interfacial points and with $\bar{u}$ denoting the average value of u constant in time.

For any $\epsilon > 0$, the energy

$$E^\epsilon[u] = \frac{1}{L}\int_0^L \left(\frac{\epsilon}{2}|u'|^2 + \frac{1}{\epsilon}F(u) + \frac{1}{2}(v')^2\right) dx \tag{79}$$

is a Lyapunov functional [5] for the original model problem (1)–(3). (Recall that v is given in terms of u by (3).)

The local minima of (79) were considered by Ren and Wei [9]; they showed that they are approximated by the local minima of the Gamma-limit functional

$$E^0[N, \xi] = \frac{\mu N}{L} + E_*[N, \xi], \tag{80}$$

where $\mu = \int_{-1}^1 \sqrt{2F(u)}du$, $\xi = (\xi_1, \dots, \xi_N)$, and

$$E_*[N, \xi] = \frac{1}{L}\int_0^L \frac{1}{2}(v'(x))^2 dx. \tag{81}$$

The domain of (79) consists of L^2 functions $u(x)$ on $(0, L)$ with prescribed average $\bar{u}$; the value of the functional may be $+\infty$. The reduced energy (81) can also be interpreted as a functional of functions u, but we write it as depending on N and the ξ_n. It is calculated as follows. Let ξ_n be given. Setting $p_0 = 0$, one calculates in succession the numbers p_n from (21) and then the function $v'(x)$ from (18). The fact that $\bar{u}$ is the average of u guarantees that $v' = 0$ at $x = 0, L$. This function is then used in the definition (81).

It can also be shown that in our 1D scenario, $E_*[N, \xi]$ is a Lyapunov functional for the FBP (12)–(17). Since the first part of (80) depends only on N and N doesn't change, $E^0[N, \xi]$ is also a Lyapunov functional.

If one allows N to be variable, then the global minimum of $E^0[N, \xi]$ is attained when $N = N_m$, a value which depends only on L and can be calculated, and when the numbers ξ_n, $n = 0, \dots, N_m$ form a stationary solution (section 6).

Calculations show (see e.g. [9]) that N_m is an integer differing from

$$\left(\frac{1}{24\mu}\right)^{1/3} L \tag{82}$$

by no more than unity.

The spacing which minimizes this energy is then approximately $(24\mu)^{1/3}$, which is independent of L.

However, this global minimum has little relevance for the 1D dynamical FBP because N remains constant. The artificial nucleation process described in section 9 allows for only an increase, not a decrease, in N. It is likely that when $\epsilon \ll 1$, layered solutions of the original PDEs (1)–(3) follow the evolving solution of the FBP, so that the global minimum will not be very important for their evolution either.

There remains the question as to whether solutions of the original problem that begin without layers will naturally develop layers with the preferred spacing. In this connection it is important to look at the dominant modes arising from initial data which are a small perturbation of the constant solution $u \equiv \bar{u}$, according to the evolution (1)–(3). It turns out that the dominant modes have wavelength of the order $O(\epsilon)$, rather than $O(1)$, which is characteristic of the optimal spacing.

11. Discussion. The energy-conserving, mass-preserving (if u is interpreted as a scaled mass density) pattern-forming system of PDEs (1)–(3) proposed by Bahiana and Oono and by Nishiura and Ohnishi has, as the latter authors showed, a formal limiting FBP, and this connection was justified in a weak sense for radial solutions by Henry. In the 1D case, this latter problem is shown to reduce to a discretized heat equation. These equations have an explicit solution. In fact, we have the representation (45)–(48) (or (33) if $\bar{u} = 0$) for the solution in terms of its initial data.

Generally, it is shown that stationary configurations (in which the intervals where $u = 1$ all have the same length, and the same is true where $u = -1$) have a large basin of stability. When, for example, the initial data $\xi_n(0)$ approach a stationary solution as $|n| \to \infty$, the evolution approaches that same configuration uniformly in n as $t \to \infty$.

The system (35), (36), being quasi-monotone, enjoys a maximum principle, which we have in effect used in Lemma 1. It can be used to prove other properties of the solutions not considered here, such as comparison principles.

By the nature of the FBP, interfaces in one dimension are neither created nor destroyed. (However, an artificial nucleation process was described in section 9.) Therefore, when the domain is a finite interval, stability is of course within the class of evolutions with a given number of interfaces. Ren and Wei [9] considered, for the original problem ($\epsilon > 0$) in one dimension, the question of which configurations minimize the energy E^ϵ (79). Here the minimization is over all functions, so in a sense the number of interfaces (which appear as transition layers in this case) can be varied. They obtained an "optimal" spacing of the order $O(1)$ with the spacing given in accordance with (82). In [8, 10], the same analysis was done for a free energy corresponding to the alternative evolution (1), (4), (3). Similar results were obtained, but their optimal spacing was $O(\hat{\epsilon}^{1/3})$ (as $\hat{\epsilon} \to 0$). If we invoke the scaling procedure connecting their energy functional with ours, their spacing of order $O(\hat{\epsilon}^{1/3})$ corresponds to our spacing of $O(1)$ as $\epsilon \to 0$.

The role of global (as opposed to local) minimizers is much more important for problems in higher dimensions than in one dimension, because in one dimension there is no mechanism for changing the number of interfaces once that number has been established by the initial data. Therefore, there is no mechanism for adjusting the spacing to achieve energy optimality.

REFERENCES

[1] M. Bahiana and Y. Oono, *Cell dynamical system approach to block copolymers*, Phys. Rev. A (3), 41 (1990), pp. 6763–6771.

[2] M. Henry, *Singular limit of a fourth order problem arising in the micro-phase separation of diblock copolymers*, Adv. Differential Equations, to appear.

[3] K. Kawasaki, T. Ohta, and M. Kohrogui, *Equilibrium morphology of block copolymer melts.* 2, Macromolecules, 21 (1988), pp. 2972–2980.

[4] L. Leibler, *Theory of microphase separation in block copolymers*, Macromolecules, 13 (1980), pp. 1602–1617.

[5] Y. Nishiura and I. Ohnishi, *Some mathematical aspects of the micro-phase separation in diblock copolymers*, Phys. D, 84 (1984), pp. 31–39.

[6] I. Ohnishi and Y. Nishiura, *Spectral comparison between the second and fourth order equations of conservative type with non-local terms*, Japan J. Indust. Appl. Math., 15 (1998), pp. 253–262.

[7] T. Ohta and K. Kawasaki, *Equilibrium morphology of block copolymer melts*, Macromolecules, 19 (1986), pp. 2621–2632.

[8] I. Ohnishi, Y. Nishiura, M. Imai, and Y. Matsushita, *Analytical solutions describing the phase separation driven by a free energy functional containing a long-range interaction term*, Chaos, 9 (1999), pp. 329–341.

[9] X. Ren and J. Wei, *On the multiplicity of solutions of two nonlocal variational problems*, SIAM J. Math. Anal., 31 (2000), pp. 909–924.

[10] X. Ren and J. Wei, *On Energy Minimizers of the Di-Block Copolymer Problem*, preprint, Utah State University, Logan, UT, 2001.

SIAM J. MATH. ANAL.
Vol. 33, No. 3, pp. 607–633

DIFFUSIVE N-WAVES AND METASTABILITY IN THE BURGERS EQUATION[*]

YONG JUNG KIM[†] AND ATHANASIOS E. TZAVARAS[‡]

Abstract. We study the effect of viscosity on the large time behavior of the viscous Burgers equation by using a transformed version of Burgers (in self-similar variables) that captures efficiently the mechanism of transition to the asymptotic states and allows us to estimate the time of evolution from an N-wave to the final stage of a diffusion wave. Then we construct certain special solutions of diffusive N-waves with unequal masses. Finally, using a set of similarity variables and a variant of the Cole–Hopf transformation, we obtain an integrated Fokker–Planck equation. The latter is solvable and provides an explicit solution of the viscous Burgers equation in a series of Hermite polynomials. This format captures the long-time–small-viscosity interplay, as the diffusion wave and the diffusive N-waves correspond, respectively, to the first two terms in the Hermite polynomial expansion.

Key words. diffusion waves, diffusive N-waves, convection-diffusion, metastability

AMS subject classifications. 76N17, 35L65, 76Rxx

PII. S0036141000380516

1. Introduction.

The Cauchy problem for the viscous Burgers equation

$$\begin{gathered} u_t + uu_x = \mu u_{xx}, \qquad x \in \mathbb{R}\,,\ \mu, t > 0, \\ u(x,0) = u_0(x), \qquad x \in \mathbb{R}, \end{gathered} \tag{1.1}$$

has, since the pioneering work of Hopf [7], served as a paradigm for the development of the theory of shock waves (see [4] and references therein).

In the limit as the viscosity $\mu \to 0$, the solution u^μ of (1.1) converges to the entropy weak solution u of the inviscid Burgers

$$\begin{gathered} u_t + uu_x = 0, \qquad x \in \mathbb{R}\,,\ t > 0, \\ u(x,0) = u_0(x), \quad x \in \mathbb{R}, \end{gathered} \tag{1.2}$$

satisfying the Oleinik condition [16]

$$u(x+,t) \le u(x-,t)\,.$$

The asymptotic behavior of (1.2) is an N-wave, whose positive and negative masses are determined by the positive and negative invariants

$$\begin{gathered} p_0 = -\inf_x \int_{-\infty}^x u_0 dy\,, \quad q_0 = \sup_x \int_x^\infty u_0 dy\,, \\ \text{where} \quad -p_0 + q_0 = M_0 = \int_{\mathbb{R}} u_0(x)dx, \end{gathered} \tag{1.3}$$

[*] Received by the editors November 6, 2000; accepted for publication May 3, 2001; published electronically October 4, 2001. This research was partially supported by the National Science Foundation.
http://www.siam.org/journals/sima/33-3/38051.html

[†] Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, MN 55455-0436 (yjkim@ima.umn.edu).

[‡] Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53706 (tzavaras@math.wisc.edu).

of the initial data [7, 12]. On the other hand, the large time behavior of u^μ for fixed μ is characterized by the well-known diffusion wave of mass M_0 (see [7]). Therefore, the reversal of order in the successive limit passages $t \to \infty$, $\mu \to 0$ leads to different results; in other words, the long-time response of the viscous Burgers equation exhibits sensitive dependence on the viscosity.

The objective of the present article is to provide a quantitative understanding of the long-time–small-viscosity interplay for the Burgers equation. To place the problem in context, the reader is referred to the numerical runs of section 4 for the evolution of solutions to (1.1), when the viscosity $\mu \ll 1$. These indicate that at an initial stage u^μ evolves from the initial state u_0 into a saw-tooth profile; at a second stage, the waves interact eventually producing an approximate N-wave; this last structure persists for a very long time, but eventually the smallest of the positive and negative masses is consumed, and thereafter u^μ looks like the final asymptotic state of a diffusion wave.

The Burgers equation is invariant under the group of scaling transformations $x \to cx$, $t \to c^2 t$, $u \to u/c$, and under time and space translations, $t \to t + a$ and $x \to x + b$. This property suggests a transformation to similarity variables,

$$s = \ln(t)\,, \quad \xi = x/\sqrt{t}\,, \quad w(\xi, s) = \sqrt{t}\, u(x,t)\,, \tag{1.4}$$

which puts (1.1) in the form

$$w_s + \left(\frac{1}{2}w^2 - \frac{1}{2}\xi w\right)_\xi = \mu w_{\xi\xi}, \qquad \xi\,, s \in \mathbb{R}, \quad \mu > 0\,. \tag{1.5}$$

The diffusion waves (see (2.6)) are the steady states of (1.5) and determine its large time behavior (see [7] and sections 2 and 3). In the limit $\mu \to 0$, solutions of (1.5) satisfy the (self-similar variant of the) inviscid Burgers equation

$$w_s + \left(\frac{1}{2}w^2 - \frac{1}{2}\xi w\right)_\xi = 0, \tag{1.6}$$

subject to the Oleinik entropy condition

$$w(\xi+, s) \le w(\xi-, s). \tag{1.7}$$

The admissible steady states of (1.6)–(1.7) are the two parameter family

$$\mathcal{N}_{p,q}(\xi) = \begin{cases} \xi, & -\sqrt{2p} < \xi < \sqrt{2q}, \\ 0, & \text{otherwise,} \end{cases}$$

with p, q positive constants. They are precisely the self-similar form of the N-waves, and p, q measure, respectively, the negative and positive mass of the N-wave.

The similarity forms (1.5) and (1.6)–(1.7) provide a convenient formulation for performing long-time numerical runs as well as a framework for a qualitative explanation of the various regimes of the problem. In the first stage the solution evolves from u_0 to a saw-tooth profile via the usual compression-attenuation mechanism of hyperbolic equations. In the next stage the waves interact and produce an approximate N-wave. In both of these stages the effect of diffusion lies only in selecting admissible discontinuities, and the evolution is essentially governed by the convection equation (1.6)–(1.7). The N-waves are steady states for (1.6), but, due to the presence of small diffusion, they are only approximate solutions for (1.5). This discrepancy drives the evolution in the last stage from an approximate N-wave to the steady state of (1.5), a diffusion wave. This stage is a very slow transition and a manifestation of metastability driven by the small diffusion.

A number of techniques have been developed for treating asymptotic behavior problems for nonlinear convection (see, e.g., [2, 3, 13]) or convection-diffusion equations (see, e.g., [5, 6, 14]). Here we exploit the similarity structure based on the invariances of the Burgers equation. This perspective is initiated in Hopf [7] and is developed in Tartar [20], Liu–Pierre [13] (for convection equations), and Escobedo–Vazques–Zuazua [5, 6] (for various multidimensional convection-diffusion equations with power laws). The aim is to pursue quantitative explanations of the various regimes; the simplicity of the equation allows us to obtain a complete picture, including the interesting regime of small-viscosity long-time interaction.

We begin, in sections 2–3, with a review of some of the standard hyperbolic theory of L^1-contraction and Oleinik inequality for the self-similar Burgers (1.5). We show how to use a special Lyapunov function to establish the asymptotic in time profile (for μ fixed) of a diffusion wave (see Theorem 3.3). The results in sections 2–3 can be established with different methods, but we draw attention to the fact that the study of (1.5) provides optimal convergence results with little effort. They also set the stage for section 4, where we provide explanations and predictions for the numerical runs based on heuristic arguments and theoretical results.

In the last two sections we turn our attention to the issue of metastability. In section 5, we generalize a construction of Whitham [21] and provide a special solution corresponding to a diffusive N-wave with unequal positive and negative masses. This solution reads

$$\begin{aligned} &u_{p,q}(x,t+1) = \sqrt{\frac{\mu}{t+1}}\, v_{p,q}\left(\frac{x}{\sqrt{4\mu(t+1)}}, t\right), \\ &\qquad \text{where} \quad v_{p,q}(\xi,t) = \frac{\frac{B}{\sqrt{\pi}} e^{-\xi^2} + 2A\frac{1}{\sqrt{t+1}}\xi e^{-\xi^2}}{1 - \frac{B}{\sqrt{\pi}}\int_{-\infty}^{\xi} e^{-\zeta^2} d\zeta + A\frac{1}{\sqrt{t+1}} e^{-\xi^2}}, \end{aligned} \tag{1.8}$$

where $B = 1 - e^{(p-q)/2\mu}$ and $A = e^{p/2\mu} + O(B)$ are constants determined via the positive and negative masses p, q of the initial data. The solution $u_{p,q}$ converges to an N-wave as $\mu \to 0$, which suggests the terminology diffusive N-wave for $u_{p,q}$. It captures, in an explicit manner, the slow transition from an approximate N-wave to the final stage of a diffusion wave.

Motivated by the format of (1.8), we use in section 6 appropriate similarity variables and a variant of the Cole–Hopf transformation to transform the Burgers equation into an integrated version of a Fokker–Planck equation. Unlike the Laplace operator on the real line, the present operator (6.7) has a discrete spectrum. The process yields an explicit formula for the solution of the Burgers equations in terms of Hermite polynomials:

$$\begin{aligned} &u(x,t) = \sqrt{\frac{\mu}{t+1}}\, v\left(\frac{x}{\sqrt{4\mu(t+1)}}, t\right), \\ &\qquad \text{where} \quad v(\xi,t) = -\frac{\partial_\xi \psi_M^\infty + \sum_{n=0}^{\infty} a_n (t+1)^{-\frac{n+1}{2}} \partial_\xi\left(H_n(\xi) e^{-\xi^2}\right)}{\psi_M^\infty + \sum_{n=0}^{\infty} a_n (t+1)^{-\frac{n+1}{2}} H_n(\xi) e^{-\xi^2}}, \end{aligned} \tag{1.9}$$

where ψ_M^∞ is the potential of a diffusion wave of mass M (see (6.16)), H_n are the Hermite polynomials, and the Fourier–Hermite coefficients a_n are computed explicitly from the data (see (6.18)). The asymptotic profile of the diffusion wave corresponds to keeping only the first term in the expansion, $-\frac{\partial_\xi \psi_M^\infty}{\psi_M^\infty}$, while the diffusive N-wave (1.8) corresponds to keeping the first two terms in the expansion and describes the first order correction of the general solution beyond the diffusion wave.

2. Self-similar Burgers—preliminaries.

Some of the basic properties of the viscous Burgers equations are directly linked to the self-similar Burgers equation

$$w_s + \left(\frac{1}{2}w^2 - \frac{1}{2}\xi w\right)_\xi = \mu w_{\xi\xi}. \tag{2.1a}$$

This problem is obtained from (1.1) in two different ways. If the similarity transformation (1.4) is used, then one obtains (2.1a) set in the interval $(\xi, s) \in \mathbb{R} \times \mathbb{R}$, and the initial data are projected back to $-\infty$. Alternatively, and due to the invariance $t \to t+1$, one may use the transformation

$$s = \ln(t+1)\,, \quad \xi = x/\sqrt{t+1}\,, \quad w(\xi, s) = \sqrt{t+1}\,u(x,t)$$

and obtain the initial value problem consisting of (2.1a) set on $(\xi, s) \in \mathbb{R}\times\mathbb{R}^+$ subject to data

$$w(\xi, 0) = w_0(\xi). \tag{2.1b}$$

In either case the total mass remains invariant under the transformation, and the initial mass is preserved:

$$\int_{\mathbb{R}} w(\xi, s) d\xi = M_0 < \infty.$$

Let w^μ be the solution of (2.1). In the limit $\mu \to 0$, $w^\mu \to w$ a.e., where w is a weak solution of the initial value problem

$$\begin{aligned} w_s + \left(\frac{1}{2}w^2 - \frac{1}{2}\xi w\right)_\xi &= 0, \\ w(\xi, 0) &= w_0(\xi), \end{aligned} \tag{2.2}$$

that satisfies the Oleinik (entropy) condition

$$w(\xi+, s) \le w(\xi-, s). \tag{2.3}$$

2.1. Steady states.

The admissible steady states of (2.2) are given by the two parameter family

$$\mathcal{N}_{p,q}(\xi) = \begin{cases} \xi, & -\sqrt{2p} < \xi < \sqrt{2q}, \\ 0, & \text{otherwise}, \end{cases} \tag{2.4}$$

parametrized by the positive constants p and q measuring, respectively, the mass of the negative and positive parts of the steady state. These solutions are precisely the N-waves, when viewed on the self-similar coordinates, and are denoted by $\mathcal{N}_{p,q}$. If the total mass M is prescribed, there is a one parameter family $\mathcal{N}_{p,p+M}$ corresponding to the mass M.

The equilibria of (2.1a) (for $\mu > 0$) satisfy the equation

$$\left(\frac{1}{2}\mathcal{G}^2 - \frac{1}{2}\xi\mathcal{G}\right)' = \mu\mathcal{G}'' \tag{2.5}$$

and are computed by the formula

$$\mathcal{G}_M(\xi) = \frac{\sqrt{\mu}\,(1 - e^{-M/2\mu})\,e^{-\xi^2/4\mu}}{1 - (1 - e^{-M/2\mu})\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\xi/\sqrt{4\mu}} e^{-\zeta^2}d\zeta}, \qquad M \in \mathbb{R}. \tag{2.6}$$

The denominator in (2.6) does not vanish, and the total mass of $\mathcal{G}_M$ is computed by

$$\int_{-\infty}^{\infty} \mathcal{G}_M d\xi = -2\mu \int_{-\infty}^{\infty} \partial_\xi \ln\left(1 - (1 - e^{-M/2\mu})\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\xi/\sqrt{4\mu}} e^{-\zeta^2}d\zeta\right) d\xi = M. \tag{2.7}$$

The steady states of the viscous problem are thus determined by the total mass M and correspond to the well-known diffusion waves in self-similar coordinates. We summarize some of their properties below.

Lemma 2.1. *Let $\mathcal{G}_M$ be a diffusion wave given by* (2.6). *Then*

(i) $\mathcal{G}_0(\xi) = 0$ *and* $\int_{-\infty}^{\infty} \mathcal{G}_M d\xi = M$*;*

(ii) $\mathcal{G}_M(\xi) > \mathcal{G}_{M'}(\xi)$ *for* $M > M'$*;*

(iii) *for any bounded function $f(\xi)$ with a compact support, there exists $M > 0$ such that $\mathcal{G}_{-M}(\xi) \leq f(\xi) \leq \mathcal{G}_M(\xi)$;*

(iv) $\mathcal{G}_M \to \mathcal{N}_{0,M}$ *as $\mu \to 0$ for $M > 0$ and $\mathcal{G}_M \to \mathcal{N}_{|M|,0}$ as $\mu \to 0$ for $M < 0$.*

2.2. Oleinik inequality. We present a quick derivation of some well-known properties viewed from the perspective of (2.1a). Recall that (2.1a) arises from the transformation (1.4) and is set on $\mathbb{R} \times \mathbb{R}$. The derived estimates are independent of μ and s. We begin with the analogue of the Oleinik estimate [16].

Lemma 2.2. *The solution $w(\xi, s)$ of* (2.1a) *satisfies*

$$w_\xi(\xi, s) \leq 1, \qquad s, \xi \in \mathbb{R}. \tag{2.8}$$

Proof. The quantity $z = w_\xi$ satisfies

$$z_s + \left(w - \frac{1}{2}\xi\right) z_\xi + z(z-1) = \mu z_{\xi\xi}. \tag{2.9}$$

If z has an interior maximum, then the value at the maximum is between 0 and 1. Since

$$w_\xi(\xi, \ln t) = t u_x(\sqrt{t}\xi, t)\,, \tag{2.10}$$

for smooth data u_0 we have $\lim_{s\to-\infty} w_\xi(\xi, s) = 0$, and (2.8) follows. If the data are not smooth, they may be approximated by smooth data in a standard way, and we conclude (2.8) by a density argument. □

Consider the functions

$$\begin{aligned} W(\xi, s) = W_-(\xi, s) &= \int_{-\infty}^{\xi} w(\zeta, s)d\zeta, \\ W_+(\xi, s) &= \int_{\xi}^{\infty} w(\zeta, s)d\zeta = M_0 - W(\xi, s) \end{aligned} \tag{2.11}$$

and the quantities

$$(2.12)\qquad \begin{aligned} p(s) &= -\inf_{\xi} W(\xi,s)\,,\\ q(s) &= \sup_{\xi} W_+(\xi,s) = M_0 + p(s)\,. \end{aligned}$$

$p(s)$ and $q(s)$ are time-invariants for solutions of the inviscid problem [12], but for the viscous problem they do not remain constant anymore. W satisfies a viscous Hamilton–Jacobi equation

$$(2.13)\qquad W_s + \frac{1}{2}(W_\xi - \xi)W_\xi = \mu W_{\xi\xi}.$$

As a simple implication of the maximum principle and Lemma 2.2, we have the following lemma.

Lemma 2.3. *Let $w(\xi,s;\mu)$ be the solution of* (2.1a), *let W be its integral given by* (2.11), *and let $A = -\inf W(\xi,0)$ and $B = \sup W(\xi,0) \geq 0$. Then W and w are uniformly bounded by*

$$(2.14)\qquad -A \leq W(\xi,s) \leq B,$$

$$(2.15)\qquad |w(\xi,s)| \leq \sqrt{2(A+B)}.$$

Proof. The estimate (2.14) follows from the maximum principle. To show (2.15), suppose that $w(\xi_1,s) > \sqrt{2(A+B)}$ for some s,ξ_1. Let $\xi_0 < \xi_1$ be such that $w(\xi_0) = 0$ and $w(\xi,s) > 0$ on (ξ_0,ξ_1). (If $w > 0$ on $(-\infty,\xi_1)$, we take $\xi_0 = -\infty$.) Since $w_\xi \leq 1$, we have $\int_{\xi_0}^{\xi_1} w(\xi,s)d\xi > A+B$ and

$$W(\xi_1,s) = W(\xi_0,s) + \int_{\xi_0}^{\xi_1} w(\xi,s)d\xi > B,$$

thus violating (2.15). If it is assumed that $w(\xi_1,s) < -\sqrt{2(A+B)}$, then similar arguments lead to a contradiction. □

2.3. L^1-contraction theory. From now on, consider the initial value problem (2.1) consisting of (2.1a) and (2.1b) and set $(\xi,s) \in \mathbb{R}\times\mathbb{R}^+$. For data $w_0 \in L^\infty \cap L^1$ this problem has a unique smooth solution. Let w_1, w_2 be two solutions; their difference $v = w_1 - w_2$ satisfies the linear parabolic equation

$$(2.16)\qquad \begin{aligned} v_s + \frac{1}{2}((w_1+w_2-\xi)v)_\xi &= \mu v_{\xi\xi},\\ v(\xi,0) &= w_1(\xi,0) - w_2(\xi,0). \end{aligned}$$

The integral

$$V(\xi,s) = \int_{-\infty}^{\xi} v(\zeta,s)d\zeta = \int_{-\infty}^{\xi} w_1(\zeta,s) - w_2(\zeta,s)d\zeta$$

satisfies

$$(2.17)\qquad V_s + \frac{1}{2}(w_1+w_2-\xi)V_\xi - \mu V_{\xi\xi} = 0.$$

Let $v^+ = \max\{v, 0\} = (v + |v|)/2$.

THEOREM 2.4. *Let* w_1, w_2 *be classical solutions of* (2.1) *with initial data* $w_1(\xi, 0)$ *and* $w_2(\xi, 0)$*. Then*

$$\int_{-\infty}^{\infty} |w_1(\xi, s) - w_2(\xi, s)| d\xi \le \int_{-\infty}^{\infty} |w_1(\xi, 0) - w_2(\xi, 0)| d\xi, \tag{2.18}$$

$$\int_{-\infty}^{\infty} (w_1(\xi, s) - w_2(\xi, s))^{\pm} d\xi \le \int_{-\infty}^{\infty} (w_1(\xi, 0) - w_2(\xi, 0))^{\pm} d\xi. \tag{2.19}$$

Furthermore, the quantity

$$\int_{-\infty}^{\infty} (w_1(\xi, s) - w_2(\xi, s))^{\pm} d\xi \tag{2.20}$$

decreases strictly in s*, unless either* $w_1(\xi, s) \le w_2(\xi, s)$ *or* $w_2(\xi, s) \le w_1(\xi, s)$ *for all* $\xi \in \mathbb{R}$.

Proof. The first part is a direct consequence of the standard contraction theory for convection-diffusion equations; see [8, 18]. We present a proof of the second part, which is based on a detailed analysis of a linear equation

$$v_s + (a(\xi, s)v)_\xi = \mu v_{\xi\xi}, \tag{2.21}$$

with $a(\xi, s)$ continuously differentiable and $|a_\xi| \le M$.

Step 1. Let $v^+ = \max\{v, 0\} = \frac{v+|v|}{2}$. Then $\partial_\xi v^+ = v_\xi \chi_{\{v>0\}}$, $\partial_s v^+ = v_s \chi_{\{v>0\}}$, and $v^+ \in W^{1,\infty}(\mathbb{R} \times \mathbb{R}^+)$. Also set

$$V^+(\xi, s) = \int_{-\infty}^{\xi} v^+(\zeta, s) d\zeta = \int_{-\infty}^{\xi} (w_1 - w_2)^+(\zeta, s) d\zeta, \tag{2.22}$$

which (due to the integrability of v and v_s) enjoys the regularity $W^{2,\infty}$ in ξ and $W^{1,\infty}$ in s. We show in this step that v^+ and V^+ satisfy

$$(v^+)_s + (a(\xi, s)v^+)_\xi \le \mu(v^+)_{\xi\xi}, \tag{2.23}$$

$$(V^+)_s + a(\xi, s)(V^+)_\xi \le \mu(V^+)_{\xi\xi} \tag{2.24}$$

in the sense of distributions.

To see that, consider the functions

$$\psi_n(v) = \begin{cases} 0, & v \le 0, \\ nv, & 0 \le v \le 1/n, \\ 1, & v \ge 1/n, \end{cases} \qquad \Psi_n(v) = \int_{-\infty}^{v} \psi_n(\tau) d\tau = \begin{cases} 0, & v \le 0, \\ \dfrac{nv^2}{2}, & 0 \le v \le 1/n, \\ v - \dfrac{1}{2n}, & v \ge 1/n. \end{cases}$$

Then $\Psi_n(v)$ satisfies

$$\begin{aligned} (\Psi_n(v))_s + (a\Psi_n(v))_\xi + a_\xi(v\psi_n(v) - \Psi_n(v)) &= \mu(\Psi_n(v))_{\xi\xi} - \mu\psi_n'(v)v_\xi^2 \\ &\le \mu(\Psi_n(v))_{\xi\xi}. \end{aligned} \tag{2.25}$$

Using the properties

$$\Psi_n(v) \to v^+, \quad 0 \le \Psi_n(v) \le v^+, \tag{2.26}$$

$$v\psi_n(v) - \Psi_n(v) \to 0\,, \quad |v\psi_n(v) - \Psi_n(v)| \le \frac{1}{2}v^+, \tag{2.27}$$

we pass to the limit $n \to \infty$ in (2.25) and obtain (2.23).

Next we consider the integrated version of (2.25),

$$\begin{aligned} \partial_s \int_{-\infty}^{\xi} \Psi_n(v) d\zeta + a\Psi_n(v) + \int_{-\infty}^{\xi} a_\xi (v\psi_n(v) - \Psi_n(v)) d\zeta \\ = \mu \partial_\xi \Psi_n(v) - \mu \int_{-\infty}^{\xi} \psi_n'(v) v_\xi^2 d\zeta \le \mu \partial_\xi \Psi_n(v), \end{aligned} \tag{2.28}$$

and use the properties (2.25), $|a_\xi| \le M$, and

$$\int_{-\infty}^{\xi} \Psi_n(v) d\zeta \to \int_{-\infty}^{\xi} v^+ d\zeta\,, \quad \int_{-\infty}^{\xi} \Psi_n(v) d\zeta \le V^+$$

$$\left| \int_{-\infty}^{\xi} a_\xi (v\psi_n(v) - \Psi_n(v)) d\zeta \right| \le \frac{M}{2} \int_{-\infty}^{\xi} v^+ d\zeta$$

to pass to the limit $n \to \infty$ and derive (2.24).

Step 2. From (2.23) we derive the inequality

$$\frac{d}{ds} \int \varphi(\xi) v^+(\xi, s) d\xi - \int \big(a v^+ - \mu (v^+)_\xi \big)(\xi, s) \varphi_\xi(\xi) d\xi \le 0 \tag{2.29}$$

for any positive test function $\varphi \in C_c^\infty(\mathbb{R})$. In turn, this yields that $\int v^+(\xi, s) d\xi$ is decreasing in s. Since $\int v(\xi, s) d\xi$ is a conserved quantity, this implies (2.18) and (2.19).

We may obtain a more detailed variant of (2.19) as follows. Fix $s > 0$ and consider a decomposition of the open set $\{\xi : v(\xi, s) > 0\} = \cup_k (\alpha_k, \beta_k)$ into countably many subintervals such that $v(\cdot, s) > 0$, and C^1 on (α_k, β_k), $v_\xi(\alpha_k, s) \ge 0$, and $v_\xi(\beta_k, s) \le 0$. Consider a test function φ_n such that

$$\varphi_n(\xi) = \begin{cases} 0, & \xi \le \alpha_k \ \text{ or } \ \xi \ge \beta_k, \\ 1, & \alpha_k + \dfrac{1}{n} \le \xi \le \beta_k - \dfrac{1}{n}, \\ \text{linear}, & \alpha_k < \xi < \alpha_k + \dfrac{1}{n} \ \text{ or } \ \beta_k - \dfrac{1}{n} < \xi < \beta_k. \end{cases}$$

We apply (2.29) for this test function and pass to the limit $n \to \infty$ to obtain

$$\frac{d}{ds} \int_{\alpha_k}^{\beta_k} v^+(\xi, s) d\xi \le \mu \big(v_\xi(\beta_k, s) - v_\xi(\alpha_k, s) \big) \le 0\,,$$

$k = 1, 2, \dots$; that is, the area under any component of v^+ is decreasing in size.

Step 3. The second part of the theorem, which is a stronger version of the first part, is obtained from the strong maximum principle. Consider $V^+(\xi, s)$ in (2.22), fix $\bar{s} > 0$, and suppose there exists $\xi_0 < \xi_1$ such that $v(\xi_0, \bar{s}) > 0$ and $v(\xi_1, \bar{s}) < 0$. Consider a restriction of V^+ defined by

$$Z^+(\xi, s) = \int_{-\infty}^{\xi} v^+(\zeta, s) \chi_{(-\infty, \xi_1)}(\zeta)\, d\zeta = \begin{cases} V^+(\xi, s) & \text{if } \xi < \xi_1, \\ V^+(\xi_1, s) & \text{if } \xi > \xi_1, \end{cases} \quad |s - \bar{s}| < \delta,$$

where

$$\chi_{(-\infty,\xi_1)} = \begin{cases} 1, & \xi \in (-\infty, \xi_1), \\ 0, & \xi \notin (-\infty, \xi_1). \end{cases}$$

We can take $\delta > 0$ so small that Z^+ has the same regularity as V^+ and compute

$$\begin{aligned} \partial_s Z^+ + a\partial_\xi Z^+ - \mu\partial_\xi^2 Z^+ \\ &= \big(\partial_s V^+ + a\partial_\xi V^+ - \mu\partial_\xi^2 V^+\big)\chi_{(-\infty,\xi_1)} - \mu v^+(\xi_1, s)\delta(\xi - \xi_1) \\ &\quad + \partial_s V^+(\xi_1, s)\chi_{(-\infty,\xi_1)} = I_1 + I_2 + I_3 \le 0, \end{aligned} \tag{2.30}$$

where the last inequality follows from (2.24) and the properties $v^+(\xi_1, s) = 0$ (hence $I_2 = 0$) and (from step 2)

$$\frac{d}{ds}\int_{-\infty}^{\xi_1} v^+(\zeta, s)d\zeta = \frac{d}{ds}V^+(\xi_1, s) \le 0.$$

Therefore, Z^+ satisfies the parabolic inequality (2.30) in the interval $0 < s - \bar{s} < \delta$.

If $Z^+(\xi_1, s)$ is strictly decreasing at $\bar{s}$, then (2.20) follows. On the other hand, if $Z^+(\xi_1, s)$ remains constant in the small time interval, then $Z^+(\cdot, s)$ has a strictly positive maximum at the interior point $\xi = \xi_1$ for a time interval $0 < s - \bar{s} < \delta$ which contradicts the strong maximum principle (see, e.g., [11]). □

Remark 2.5 (the lap-number). For semilinear parabolic equations there exists a literature concerning the number of zeroes of a solution, sometimes called the lap-number (see, e.g., Matano [15] and Angenent [1]). Such results hinge on analysis of the linear equation

$$u_t = u_{xx} + q(x, t)u, \qquad x \in \mathbb{R}\,,\ t > 0\,, \tag{2.31}$$

when $q(x, t) \in L^\infty$ and for solutions satisfying the bound $|u(x, t)| \le Ae^{Bx^2}$. It is shown in [1] that the number of zeroes

$$\mathbb{Z}_t = \{x \in \mathbb{R} : u(x, t) = 0\}$$

becomes immediately (at time $t = 0+$) a discrete set, and thereafter the number of zeroes is decreasing in time. The basis of the last result is the property that if $u(x, t)$ and $u_x(x, t)$ vanish simultaneously at (x_0, t_0) (i.e., u has a multiple zero at (x_0, t_0)), then, roughly speaking, $u(\cdot, t)$ has more zeroes for $t < t_0$ than for $t > t_0$ (see Angenent [1, Theorem B] for the precise statement).

These results apply to various semilinear parabolic equations that can (through transformations of variables) be put in the form (2.31) [1]. They also apply to quasilinear equations when linear variants of them can be put into this framework. In particular, the viscous Burgers equation shares this property (along solutions that the wave speed $u(x, t)$ and its first derivatives are uniformly bounded). Finally, the same property is transferred to solutions $w(\xi, s)$ of (2.1) through the similarity transformations.

3. Evolution of the viscous problem. Here we study the long-time convergence of solutions to a diffusion wave.

3.1. Convergence to a diffusion wave. Consider the difference $w(\xi,s)-\mathcal{G}_M(\xi)$, and note that the L^1-contraction implies

$$\int_{-\infty}^{\infty}(w(\xi,s)-\mathcal{G}_M(\xi))^{\pm}d\xi \le \int_{-\infty}^{\infty}(w(\xi,0)-\mathcal{G}_M(\xi))^{\pm}d\xi. \tag{3.1}$$

First, we prove a technical lemma, indicating that if the solution stabilizes, then the mass of the solution stabilizes.

Lemma 3.1. *Let w be the solution of* (2.1) *emanating from initial data $w_0 \in L^1\cap L^\infty$. If along a time sequence $\{s_k\}$ we have $w(\xi,s_k)\to\bar{w}(\xi)$ a.e. ξ as $s_k\to\infty$, then*

$$\lim_{s_k\to\infty}\int_{-\infty}^{\infty} w(\zeta,s_k)d\zeta = \int_{-\infty}^{\infty}\lim_{s_k\to\infty} w(\zeta,s_k)d\zeta\,. \tag{3.2}$$

Proof. Assume first that the data satisfy

$$\mathcal{G}_{-M}(\xi)\le w_0(\xi)\le\mathcal{G}_M(\xi).$$

Then from the comparison estimate (2.19)

$$\mathcal{G}_{-M}(\xi)\le w(\xi,s)\le\mathcal{G}_M(\xi)$$

and the dominated convergence theorem implies the desired result.

Now let $w_0\in L^1\cap L^\infty$. For $\varepsilon>0$ choose $M>0$ so that

$$\int_{-\infty}^{\infty}(w_0(\zeta)-\mathcal{G}_M(\zeta))^+ + (w_0(\zeta)-\mathcal{G}_{-M}(\zeta))^- d\zeta<\varepsilon. \tag{3.3}$$

Let

$$w_M(\xi,s)=\begin{cases}\mathcal{G}_M(\xi) & \text{if } w>\mathcal{G}_M,\\ w(\xi,s) & \text{if } \mathcal{G}_{-M}<w<\mathcal{G}_M,\\ \mathcal{G}_{-M}(\xi) & \text{if } w<\mathcal{G}_{-M},\end{cases} \tag{3.4}$$

and define $\bar{w}_M(\xi)$ in a similar fashion using the limit function $\bar{w}(\xi)$ in the place of $w(\xi,s)$. Observe that $w_M(\zeta,s_k)\to\bar{w}_M(\zeta)$ and that

$$\begin{aligned}&\left|\int_{-\infty}^{\infty}\big(w(\zeta,s_k)-w_M(\zeta,s_k)\big)d\zeta\right|\\ &\qquad=\left|\int_{w>\mathcal{G}_M}\big(w(\zeta,s_k)-\mathcal{G}_M(\zeta)\big)d\zeta+\int_{w<\mathcal{G}_{-M}}\big(w(\zeta,s_k)-\mathcal{G}_{-M}(\zeta)\big)d\zeta\right|\le 2\varepsilon\end{aligned}$$

from (2.19), (3.1), and (3.3). Now $\int_{\mathbb{R}} w_M(\cdot,s_k)d\zeta\to\int_{\mathbb{R}}\bar{w}_M d\zeta$, and we conclude by (3.3)

$$\left|\lim_{s_k\to\infty}\int_{-\infty}^{\infty}w(\zeta,s_k)d\zeta-\int_{-\infty}^{\infty}\bar{w}(\zeta)d\zeta\right|\le 4\varepsilon,$$

which gives (3.2). □

In the following theorem we show that the solution of (2.1) converges to a diffusion wave preserving the initial total mass.

THEOREM 3.2 (convergence in time). *Let w be the solution to the Cauchy problem* (2.1) *with $\mu > 0$ and $w_0 \in L^1 \cap L^\infty$, $\partial_\xi w_0 \in L^2$, and total mass $\int w_0(\xi) d\xi = M_0$. Then*

$$w(\xi, s) \to \mathcal{G}_{M_0}(\xi) \quad \text{as } s \to \infty, \tag{3.5}$$

a.e. and in $L^1(\mathbb{R})$, where $\mathcal{G}_{M_0}$ is the diffusion wave given by (2.6).

Proof. Consider the quantity $\Phi = e^{-\frac{1}{2\mu}W}$ motivated by the Cole–Hopf transformation. Then Φ satisfies

$$\Phi_s - \frac{1}{2}\xi\Phi_\xi = \mu\Phi_{\xi\xi}. \tag{3.6}$$

We differentiate (3.6) with respect to s and multiply by Φ_s. After rearranging the terms, we obtain

$$\partial_s\Phi_s^2 - \frac{1}{2}\partial_\xi\big(\xi\Phi_s^2\big) + \frac{1}{2}\Phi_s^2 + 2\mu\Phi_{\xi s}^2 = \mu\partial_{\xi\xi}\Phi_s^2.$$

Therefore, the quantity

$$g(s) = \int_{\mathbb{R}} \Phi_s^2(\xi, s) d\xi = \int_{\mathbb{R}} \frac{1}{4\mu^2} e^{-\frac{1}{\mu}W} W_s^2 \, d\xi$$

satisfies the differential inequality $\frac{dg}{ds} + \frac{1}{2}g \le 0$. We conclude that

$$\int_{\mathbb{R}} e^{-\frac{1}{\mu}W} W_s^2 \, d\xi \le \left(\int_{\mathbb{R}} e^{-\frac{1}{\mu}W(\xi,0)} W_s^2(\xi, 0) \, d\xi \right) e^{-\frac{s}{2}}$$

and, from (2.14) and (2.13), that

$$\int_{\mathbb{R}} W_s^2(\xi, s) \, d\xi \le O(1) e^{-\frac{s}{2}}, \tag{3.7}$$

where $O(1)$ depends on μ and the H^1-norm of the data w_0. (The last dependence may be relaxed by using the regularizing effect of (2.1), but we will not pursue the details here.)

From (2.8) and (2.15), the function $(w(\xi, s) - \xi)$ is decreasing and for any $[a, b] \subset \mathbb{R}$ we have

$$TV_{\xi\in[a,b]} w(\cdot, s) \le \sup_\xi w_0 - \inf_\xi w_0 + 2(b - a).$$

From Helly's theorem and a diagonal argument we can extract a subsequence $s_n \to \infty$ and a function $\bar{w}$ of locally bounded variation so that

$$w(\xi, s_n) \to \bar{w}(\xi) \quad \text{as } s_n \to \infty. \tag{3.8}$$

Let $\theta(\xi)$ be a C^∞-function with compact support. From (2.13) we obtain

$$\int_{\mathbb{R}} W_s(\cdot, s)\theta d\xi + \int_{\mathbb{R}} \frac{1}{2}\big(w^2 - \xi w\big)\theta d\xi + \int_{\mathbb{R}} \mu w \theta_\xi d\xi = 0.$$

We use (3.7), (3.8), and Lemma 3.1 to pass to the limit along s_n and deduce that $\bar{w}$ satisfies

$$\begin{gathered} \frac{1}{2}\bar{w}^2 - \frac{1}{2}\xi\bar{w} = \mu\bar{w}_\xi, \\ \int_{\mathbb{R}} \bar{w} d\xi = \int_{\mathbb{R}} w_0 d\xi = M_0 \, . \end{gathered} \tag{3.9}$$

Therefore, $\bar{w} = \mathcal{G}_{M_0}$, and, as the limit is uniquely determined by (3.9), we conclude that the family $w(\xi, s) \to \mathcal{G}_{M_0}(\xi)$ as $s \to \infty$. □

3.2. Convergence to a diffusion wave via the invariance principle. Next we provide a proof of the time-convergence to a diffusion wave from the viewpoint of the LaSalle invariance principle, in the spirit of [3, 6]. Recall that the solution operator of the viscous Burgers defines a contraction semigroup in $L^1(\mathbb{R})$ defined by $T(s)w_0 = w(\cdot, s)$, where $w_0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ is the initial datum.

Theorem 3.3 (convergence in time). *Let w be the solution to the Cauchy problem* (2.1) *with $\mu > 0$ and $w_0 \in L^1 \cap L^\infty$ with total mass $\int w_0(\xi)d\xi = M_0$. Then*

$$w(\xi, s) \to \mathcal{G}_{M_0}(\xi) \quad \text{as } s \to \infty, \tag{3.10}$$

pointwise and in $L^p(\mathbb{R})$, $1 \le p < \infty$, where $\mathcal{G}_{M_0}$ is the diffusion wave given by (2.6).

Proof. Consider the ω-limit set

$$\omega(w_0) = \left\{ \psi \in L^1(\mathbb{R}) : \psi = \lim_{k\to\infty} T(s_k)w_0 \text{ for a subsequence } s_k \right\}. \tag{3.11}$$

From (3.8) we have a convergent subsequence $T(s_k)w_0 = w(\cdot, s_k)$ and thus $\omega(w_0)$ is not empty. Moreover, T is positively invariant; that is, if $\psi \in \omega(w_0)$, then $T(s)\psi \in \omega(w_0)$. This follows from the semigroup property and the observation

$$T(s)\psi = \lim_{k\to\infty} T(s)T(s_k)w_0 = \lim_{k\to\infty} T(s + s_k)w_0. \tag{3.12}$$

Consider the distance functional $V : L^1(\mathbb{R}) \to \mathbb{R}$ defined by

$$V(\chi) = ||\chi - \mathcal{G}_{M_0}||_{L^1}, \tag{3.13}$$

which is continuous in $L^1(\mathbb{R})$. The diffusion wave $\mathcal{G}_{M_0}$ is a special solution of (2.1). Due to the contraction theory (2.18), V satisfies

$$V(T(s)w_0) = ||w(\cdot, s) - \mathcal{G}_{M_0}||_{L^1} \le ||w_0 - \mathcal{G}_{M_0}||_{L^1} = V(w_0) \tag{3.14}$$

for any $s > 0, w_0 \in L^1(\mathbb{R})$. Since $V(T(s)w_0)$ is decreasing in time, it converges $V(T(s)w_0) \to c$ to a limit $c \ge 0$. From the continuity of V, we obtain

$$V(T(s)\psi) = V\left(\lim_{k\to\infty} T(s + s_k)w_0\right) = \lim_{k\to\infty} V(T(s + s_k)w_0) = c \tag{3.15}$$

for any $\psi \in \omega(w_0)$ and $s \ge 0$.

We shall assume there is $\psi \in \omega(w_0)$ such that $\psi \neq \mathcal{G}_{M_0}$ and establish a contradiction. For such a ψ we have $V(\psi) = c > 0$, $T(s)\psi \in \omega(w_0)$, and $V(T(s)\psi) = c$ from (3.15). Since

$$\int \psi(\xi) - \mathcal{G}_{M_0}(\xi)d\xi = 0\,,$$

there exist ξ_0, ξ_1 such that $\mathcal{G}_{M_0}(\xi_0) > \psi(\xi_0)$ and $\mathcal{G}_{M_0}(\xi_1) < \psi(\xi_1)$. Theorem 2.4 then implies that $V(T(s)\psi) < V(\psi) = c$, which contradicts (3.15). So the nonempty ω-limit set should be $\omega(w_0) = \{\mathcal{G}_{M_0}\}$. □

4. Diffusion driven interfaces and metastability. The asymptotic behavior of the inviscid Burgers equation is an N-wave [12]. By contrast, solutions of the viscous Burgers equation approach diffusion waves asymptotically in time. That indicates a sensitive dependence of the long-time asymptotics for the viscous Burgers equation on the viscosity that we begin analyzing from this section.

Numerical computations for solutions of the viscous Burgers equation were performed using the self-similar version (1.5). This framework has two advantages: first, the solution does not spread in time (thus a smaller computational domain is needed), and second, the time variable $s = \ln(t)$ allows us to compute for long times. The results are presented in Figure 1 for a viscosity $\mu = 0.01$ (see section 4.3 for information on the numerical scheme). It is seen that initially w^{μ} evolves from "oscillatory" initial data w_0 into a saw-tooth profile. This transition occurs relatively quickly ($s \sim 0.5$) and is driven by the usual compression-attenuation mechanism of hyperbolic equations. At the next stage the waves interact and eventually produce an approximate N-wave. This takes a longer time ($s \sim 2$), and this stage is also governed by the convection mechanism mainly (see section 4.3). Once the latter stage is reached, it appears as if there is no dynamical change. A more careful look, though, shows that the negative and positive masses of the solution decrease slowly until eventually the smaller one disappears. This evolution occurs at a far slower time scale. In Figure 1 the negative mass disappears at $s \sim 100$, which, considering that the original time variable is $t = e^s$, is an exceptionally long time.

We next analyze the role of diffusion at the intermediate stage driven by wave interactions and then transition from an N-wave to a diffusion wave (for small $s \sim 0.5$ and large $s > 2$ for the numerical run of Figure 1). Then, in sections 5 and 6, we provide a quantitative description of the metastable stage of the evolution.

4.1. Wave interactions. We provide a heuristic analysis valid for solutions emanating from data that at time $s = 0$ intersect the axis at finitely many points

$$\{\xi \in \mathbb{R} : w(\xi, s) = 0\} = \{g_1(s) < g_2(s) < \cdots < g_n(s)\}, \quad s = 0, \tag{4.1}$$

and the intersections occur transversally

$$w_\xi(g_i(s), s) \neq 0, \qquad i = 1, 2, \ldots, n, \quad s = 0. \tag{4.2}$$

For notational convenience, let $g_0(s) = -\infty$ and $g_{n+1}(s) = +\infty$. (This is the generic form of solutions that appears in numerical runs after the initial stage ($s \sim 0.5$ in Figure 1); we refer to Figure 2 for such a profile.) Our goal is to track the mechanism of motion of the intersection points.

The points of intersection g_i satisfy $w(g_i(0), 0) = 0$. Using the implicit function theorem, the curves $g_i(s)$ are each defined on a maximal interval $s \in [0, S_i)$ and move with the speed

$$g_i'(s) = -\frac{w_s(g_i(s), s)}{w_\xi(g_i(s), s)}. \tag{4.3}$$

On the interval $[0, S_i)$ we have $w_\xi(g_i(s), s) > 0$, but at the maximal time S_i it is $w_\xi(g_i(S_i), S_i) = 0$. Typically, at such times two of the curves will come together and disappear.

Consider the functions

$$p_i(s) = -\int_{-\infty}^{g_i(s)} w(\zeta, s) d\zeta = -W(g_i(s), s), \tag{4.4}$$

where $-p_i(s)$ measures the total mass of $w(\cdot, s)$ to the left of $g_i(s)$. Here we define $p_i(s)$ with a negative sign since it should represent the invariance variable p_0 in (1.3)

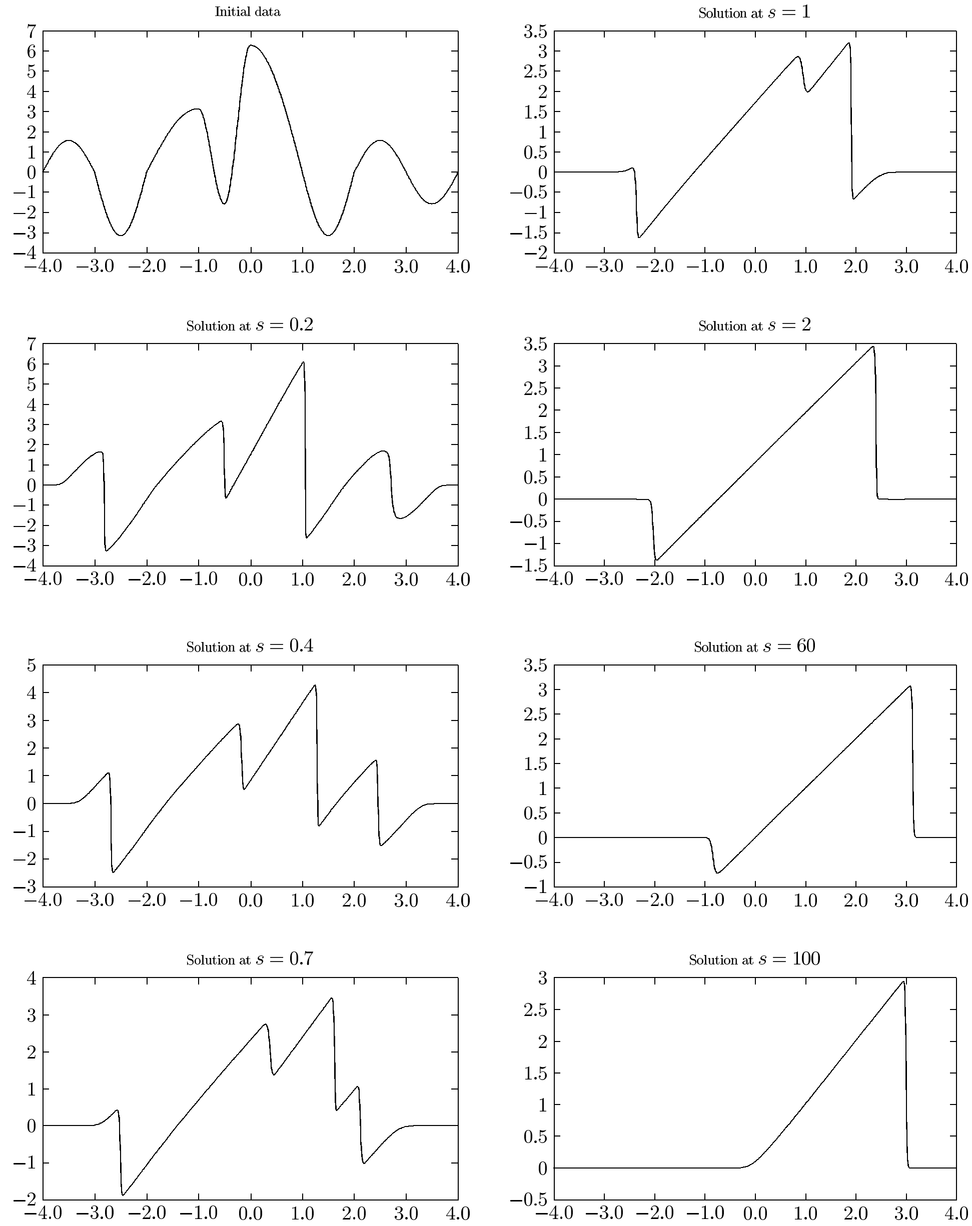


FIG. 1. *Numerical solution of transformed Burgers* (2.1): *This numerical solution is generated by the Godunov scheme with discretized viscosity and* $\Delta\xi = 0.01, \Delta s = 0.001, \mu = 0.01$.

or $p(s)$ in (2.12). Using (2.13) and the fact that $w(g_i(s), s) = 0$, we obtain

$$-p_i'(s) = W_\xi(g_i(s), s)g_i'(s) + W_s(g_i(s), s) = \mu w_\xi(g_i(s), s), \tag{4.5}$$

which pinpoints the effect of the diffusion on the mass change across a zero curve $\xi = g_i(s)$.

The remainder of our analysis is heuristic in nature. In numerical runs, after an initial transient stage, the usual compression-attenuation mechanism of Burgers

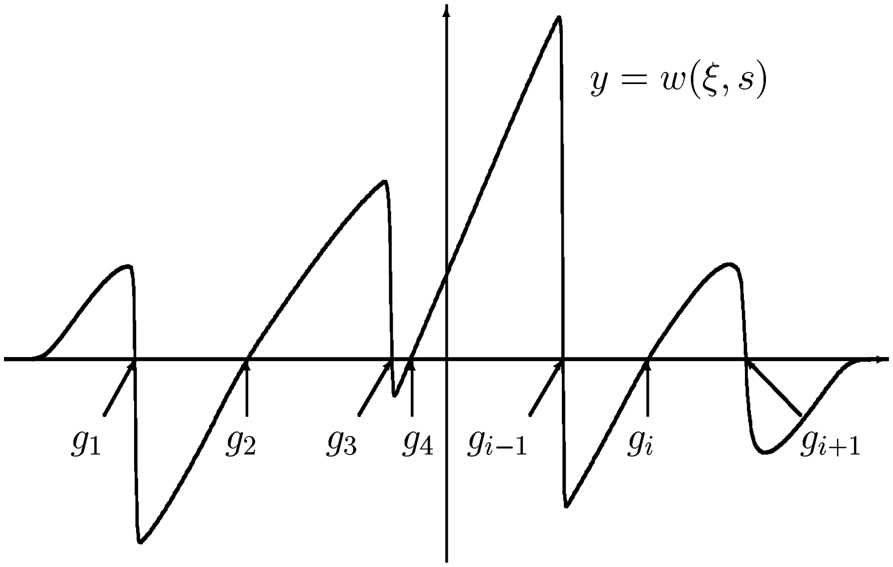


FIG. 2. *A profile of the solution at $s = 0.5$.*

produces solutions that consist entirely of shocks and rarefactions. We will operate under this condition, analyzing solutions whose time sections (s = constant) consist entirely of shocks and rarefactions and look like that in Figure 2.

The area between two zero curves $g_i(s) < g_{i+1}(s)$ (see Figure 2) is given by $A(s) = -(p_{i+1}(s) - p_i(s))$ and changes in time at a rate

$$A'(s) = -p'_{i+1}(s) + p'_i(s) = \mu w_\xi(g_{i+1}(s), s) - \mu w_\xi(g_i(s), s) < 0. \tag{4.6}$$

The positive slope at $g_i(s)$ represents an approximate rarefaction and may be estimated using the Oleinik estimate, by $0 < w_\xi(g_i(s), s) \leq 1$. Hence the change of mass across the zero curve $\xi = g_i(s)$ satisfies

$$-p'_i(s) = \mu w_\xi(g_i(s), s) \leq \mu \tag{4.7}$$

and is controlled by the diffusion. On the other hand, the negative slope at $g_{i+1}(s)$ will correspond (for μ small) to a shock profile, and the lower bound for w_ξ is of order $O(1/\mu)$. Hence the mass change across the zero curve $\xi = g_{i+1}(s)$, representing a shock, is of order $O(1)$. This points out the distinct roles of shock and rarefaction profiles: When w is increasing near a zero point, the mass change across the point is controlled by the diffusion and tends to zero as $\mu \to 0$. By contrast, near a shock profile the mass change is fast and independent of μ. The above estimations remain valid until near the time that two of the curves g_i merge and the enclosed area $A(s)$ vanishes.

4.2. Transition from N-wave to diffusion wave. Consider now a solution that emanates from N-wave like data: $w(\xi, 0) < 0$ for $\xi < \xi_0$ and $w(\xi, 0) > 0$ for $\xi > \xi_0$ with $w_\xi(\xi_0, 0) \neq 0$. We consider the zero-curve $\xi = g(s)$ emanating from the point $g(0) = \xi_0$ and satisfying $w(g(s), s) = 0$. From the implicit function theorem, the curve $g(s)$ is defined on a maximal interval $[0, S), S > 0$. As noted in Remark 2.5, the number of zeroes is nonincreasing, and for N-wave-like data, as above, it cannot happen that w and w_ξ vanish at the same point $(\bar{\xi}, \bar{s})$. This implies that either $S = \infty$ or (if S is finite) $g(s) \to \pm\infty$ as $s \to S$. In either case the solution retains its N-wave-like form in the interval $[0, S)$.

The infimum $p(s)$ of (2.12) is given by

$$p(s) = -\int_{-\infty}^{g(s)} w(\zeta, s) d\zeta. \tag{4.8}$$

From (4.7) we see that

$$p(s) \geq p(0) - \mu s \quad \text{for } 0 < s < S. \tag{4.9}$$

We next show that $S \geq \frac{1}{\mu} \min\{p(0), q(0)\}$, which provides an estimate of the transition-time to a diffusion wave. For simplicity, we assume $q(0) > p(0)$ and show that $S \geq p(0)/\mu$. Indeed, if $S < p(0)/\mu$, then from (4.9) we see that $p(s) > 0$, which implies that the mass of the negative part of the solution is still present at the time S. This contradicts the definition of S. The same argument can be made for the positive part of the solution, and we conclude that

$$\begin{array}{l} p(s) > p(0) - \mu s, \\ q(s) > q(0) - \mu s, \end{array} \qquad s < \frac{1}{\mu} \min(p(0), q(0)), \tag{4.10}$$

where $p(s), q(s)$ are the invariant variables defined in (2.12).

4.3. Comparison of the inviscid and viscous problems. We compare the evolution between the inviscid and the viscous problem for initial data consisting of two separated N-waves

$$w_0(x) = \begin{cases} x + 10, & -12 < x < -8, \\ x, & -\sqrt{2} < x < \sqrt{6}, \\ 0, & \text{otherwise}; \end{cases} \tag{4.11}$$

see Figure 3. The method of characteristics gives the exact solution of the inviscid problem: In the original variables (x, t) the solution starts to spread out, and then two inside shocks interact until a single N-wave emerges. In terms of the self-similar variables (ξ, s), one N-wave moves into the origin without changing shape until it collides with the other N-wave, and the interaction results in a new N-wave. The exact solution of the inviscid problem with initial data (4.11) is obtained by tracking the characteristics and is displayed in Figure 3 with solid lines.

The viscous problem (2.1) is solved numerically by the following scheme: Consider a uniform space $\xi_{j+1/2} = (j + 1/2)\Delta\xi$ and time $s_n = n\Delta s$ mesh, where $j \in \mathbb{Z}, n \in \mathbb{Z}^+$. The approximation of a cell-average U_j^n,

$$U_j^n \sim \frac{1}{\Delta\xi} \int_{\xi_{j-1/2}}^{\xi_{j+1/2}} w(\xi, s_n) d\xi\,,$$

is generated by a three-step explicit method,

(4.12)
$$U_j^{n+1} = U_j^n - \frac{\Delta s}{\Delta\xi}(F(U_j^n, U_{j+1}^n) - F(U_{j-1}^n, U_j^n)) + \mu \frac{\Delta s}{(\Delta\xi)^2} D(U_{j-1}^n, U_j^n, U_{j+1}^n),$$

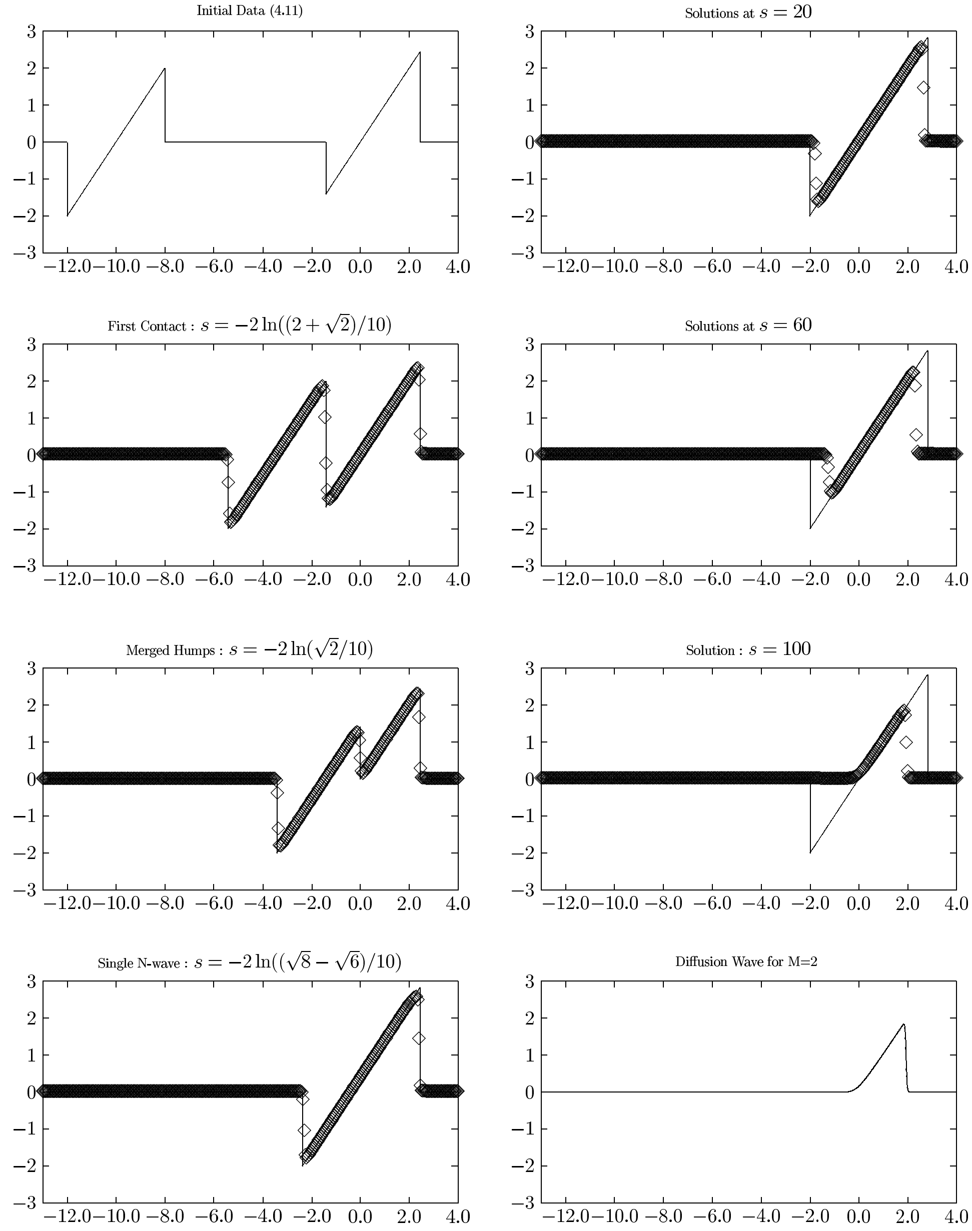


FIG. 3. *Comparison of the inviscid and viscous problems: Solid lines are exact solutions of inviscid problem* (2.2)*, and diamond dots are numerical solutions of viscous problem* (2.1) *using the Godunov scheme with* $\Delta\xi = 0.01, \Delta s = 0.0005, \mu = 0.02$*. Solutions are plotted at every other* 4 *mesh points.*

where the numerical flux F is an approximation of

$$F(U_j^n, U_{j+1}^n) \sim \frac{1}{\Delta s}\int_{s_n}^{s_{n+1}} \frac{1}{2}(w^2 - \xi_{j+1/2}w)ds, \tag{4.13}$$

and the diffusion term is discretized by

$$D(U_{j-1}^n, U_j^n, U_{j+1}^n) = U_{j-1}^n - 2U_j^n + U_{j+1}^n. \tag{4.14}$$

Since the flux of the self-similar Burgers (2.1) depends on the space variable, the solution of the Riemann problem increases exponentially along the characteristics $w(\xi(s), s) = w(\xi(0), 0)e^{s/2}$. The characteristics are not straight lines, and schemes using Riemann solver, like the Godunov scheme (see [10]), should take that into account. Here we consider a numerical flux

$$F(U_j^n, U_{j+1}^n) = \begin{cases} I(U_{j+1}^n, \bar{\xi}) & \text{if } \lambda(U_j^n, \bar{\xi}) + \lambda(U_{j+1}^n, \bar{\xi}) \le 0, \lambda(U_{j+1}^n, \bar{\xi}) \le 0, \\ I(U_j^n, \bar{\xi}) & \text{if } \lambda(U_j^n, \bar{\xi}) + \lambda(U_{j+1}^n, \bar{\xi}) > 0, \lambda(U_j^n, \bar{\xi}) > 0, \\ -3\bar{\xi}^2/8 & \text{if } \lambda(U_j^n, \bar{\xi}) < 0, \lambda(U_{j+1}^n, \bar{\xi}) > 0, \end{cases} \tag{4.15}$$

where $\bar{\xi} = \xi_{j+1/2}$ and λ is the wave speed

$$\lambda(U, \xi) = U - \xi/2. \tag{4.16}$$

$I(U, \xi)$ is an approximation of the line integral of (4.13) for a shock wave which is given by

$$I(U, \xi) = \frac{1}{2}U^2(e^{\Delta s} - 1) - \xi U(e^{\Delta s/2} - 1), \tag{4.17}$$

and we can easily check that the rarefaction wave centered at $\bar{\xi}$ has constant value $w = -3\bar{\xi}^2/8$ along the vertical line $\xi = \bar{\xi}$.

In the computation of Figure 3 the mesh size is $\Delta\xi = 0.01$, the time step is $\Delta s = 0.0005$, and the viscosity $\mu = 0.02$. The numerical solution is displayed with diamond dots. The first column indicates that the solution of the viscous problem with small viscosity is close to the solution of the inviscid problem, until the solution reaches the state of single N-wave (which is a steady state for the inviscid problem (1.6)). This stage of the evolution is dominated by convection. Subsequently, the diffusion becomes dominant, and the solution evolves slowly until it reaches the asymptotic state of a diffusion wave (a steady state for (1.5)).

Remark 4.1 (monitoring the viscosity of a numerical scheme). Numerical schemes for inviscid problems introduce numerical viscosity. A classical example is the first order Lax–Friedrich scheme for the linear equation $u_t + Au_x = 0$, which is actually a second order scheme for $u_t + Au_x = \varepsilon u_{xx}$ with a numerical viscosity

$$\varepsilon = \left(1 - \frac{(\Delta t)^2}{(\Delta x)^2}A^2\right)\frac{(\Delta x)^2}{2\Delta t}. \tag{4.18}$$

For nonlinear equations it is hard to have such an explicit control of the numerical viscosity. For that reason we opted to use a scheme based on the parabolic equation, using Godunov for the convection term and a discretization for the diffusion term.

Nevertheless, it is possible that the numerical viscosity ε can be different from the purported one μ. Under the assumption that numerical viscosity is the only factor that causes the area change of numerically approximate N-waves (an assumption clearly valid at the level of the differential equation), we can measure the numerical viscosity, using the formula (4.7), by measuring the area change and the slope at the zero point. Consider, for example, the initial data

$$w_0(\xi) = \begin{cases} x, & -2 < x < 2, \\ 0, & \text{otherwise.} \end{cases} \tag{4.19}$$

TABLE 1

Numerical viscosity of (4.12): *Initial data* (4.19), $\mu = 0.05$, $\Delta\xi = 0.01$, $\Delta s = 0.000667$.

$s^n - 1 < s < s^n$	$\max(U^n)$	$P(s^n) = \sum_{j<0} U_j^n$	$P(s^n) - P(s^n\text{-}1)$	$\Delta P/\bar{w}_\xi \sim \varepsilon$
$5 < s < 6$	1.578890	-1.686368	0.050618	0.050003
$10 < s < 11$	1.425409	-1.433271	0.050620	0.050003
$20 < s < 21$	1.068703	-0.927091	0.050614	0.050003
$35 < s < 36$	0.303368	-0.186434	0.043840	0.050083
$37 < s < 38$	0.189594	-0.110107	0.035890	0.050087
$39 < s < 40$	0.100160	-0.055559	0.024223	0.049915
$41 < s < 42$	0.045148	-0.024340	0.012980	0.049550
$43 < s < 44$	0.018336	-0.009749	0.005776	0.049215
$45 < s < 46$	0.007074	-0.003739	0.002319	0.049029
$55 < s < 56$	0.000052	-0.000027	0.000017	0.048896
$70 < s < 71$	3.135604e-08	-1.651307e-08	1.054758e-08	0.048895
$90 < s < 91$	1.812541e-12	-9.858833e-13	5.406505e-13	0.048895

As the area of the negative part is $\int_{-\infty}^{0} w(\zeta, s)d\zeta$, we consider $P(s^n) = \sum_{j<0} \Delta\xi U_j^n$. From (4.7) it is natural to define the numerical viscosity as

$$\varepsilon(s^n) = \frac{P(s^n) - P(s^{n-1})}{\Delta s\, \bar{w}_\xi}, \tag{4.20}$$

where the slope at the zero point is approximated by $\bar{w}_\xi = \frac{U_1^n - U_{-1}^n}{2\Delta\xi}$. In Table 1 we present the measured numerical viscosity for the scheme (4.12) with initial data (4.19) and viscosity $\mu = 0.05$. It is seen that the numerical viscosity represents the purported viscosity very well, $\varepsilon \sim \mu$, but as the solution decreases (in the second part of the table), the viscosity also decreases slightly. In the last part the solution is almost zero, and the numerical viscosity remains constant at the value around $\varepsilon \sim 0.049$.

5. Diffusive N-waves. The Cole–Hopf transformation implies that $u(x, t)$ is a solution of (1.1) if and only if

$$\varphi(x, t) = e^{-\frac{1}{2\mu}\int_{-\infty}^{x} u(x,t)dx} \tag{5.1}$$

solves the heat equation

$$\varphi_t = \mu\varphi_{xx}. \tag{5.2}$$

Note that (5.1) implies

$$\varphi(-\infty, t) = 1, \quad u = -2\mu\frac{\varphi_x}{\varphi}, \quad \int_a^b u(x,t)dx = -2\mu\ln(\varphi(b,t)/\varphi(a,t)) \tag{5.3}$$

and allows us to compute the mass of u. Whitham [21] uses the transformation to produce a special solution of a diffusive N-wave with equal positive and negative mass. In this section we give an extension to this construction and compute a special solution of a diffusive N-wave with possibly unequal positive and negative masses. This solution characterizes the transition from a diffusive N-wave to a diffusion wave, observed during the late-time response of Burgers.

We consider the special potential

$$\varphi_{p,q}(x, t) = 1 + A\sqrt{\frac{t_0}{t}}e^{-x^2/4\mu t} - B\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\frac{x}{\sqrt{4\mu t}}} e^{-\zeta^2}d\zeta \tag{5.4}$$

and the corresponding solution of Burgers

$$u_{p,q}(x,t) = -2\mu\frac{(\varphi_{p,q})_x}{\varphi_{p,q}} = \frac{\frac{x}{t}A\sqrt{\frac{t_0}{t}}e^{-x^2/4\mu t} + \sqrt{\frac{\mu}{t}}\frac{B}{\sqrt{\pi}}e^{-x^2/4\mu t}}{1 + A\sqrt{\frac{t_0}{t}}e^{-x^2/4\mu t} - B\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\frac{x}{\sqrt{4\mu t}}} e^{-\zeta^2}d\zeta}\,. \tag{5.5}$$

Recall that the positive and negative mass of the solution u are computed by

$$p(t) = -\inf_x \int_{-\infty}^{x} u_{p,q}(y,t)dy, \quad q(t) = \sup_x \int_x^{\infty} u_{p,q}(y,t)dy\,. \tag{5.6}$$

$p(t)$, $q(t)$ are not invariant for the viscous Burgers, but the total mass $(q(t) - p(t))$ is. The constants $A > 0$ and B in (5.4) will be determined so that the positive and negative masses at a given time t_0 are prescribed positive constants p and q, that is, $p(t_0) = p$ and $q(t_0) = q$.

This takes a lengthy computation that we outline below. To fit the total mass $M = \int u_{p,q}(y,t)dy = q - p$, we use (5.3) and obtain

$$\int_{-\infty}^{\infty} u_{p,q}(y,t)dy = -2\mu\ln(1-B) = q - p\,.$$

Hence $B = 1 - e^{(p-q)/2\mu}$. Note that $B > 0$ for $p < q$ and $B < 0$ for $q < p$; in either case $1 - B > 0$.

Clearly, $u < 0$ for $x < x_0(t) = -t\sqrt{\mu}B/\sqrt{t_0\pi}A$, and $u > 0$ for $x > x_0(t)$. The negative mass is computed from (5.3):

$$\begin{aligned} p(t) &= -\inf_x \int_{-\infty}^{x} u_{p,q}(y,t)dy = 2\mu\ln\frac{\varphi(x_0(t),t)}{\varphi(-\infty,t)} \\ &= 2\mu\ln\left(1 + A\sqrt{\frac{t_0}{t}}\,e^{-\frac{t}{t_0}(\frac{B}{\sqrt{4\pi A}})^2} - \frac{B}{\sqrt{\pi}}\int_{-\infty}^{-\sqrt{\frac{t}{t_0}}\frac{B}{\sqrt{4\pi A}}} e^{-\zeta^2}d\zeta\right). \end{aligned} \tag{5.7}$$

The requirement $p(t_0) = p$ gives the equation

$$Ae^{-(\frac{B}{\sqrt{4\pi A}})^2} = e^{p/2\mu} - 1 + \frac{B}{\sqrt{\pi}}\int_{-\infty}^{-\frac{B}{\sqrt{4\pi A}}} e^{-\zeta^2}d\zeta. \tag{5.8}$$

We give an approximate solution of (5.8). Note that $A > e^{p/2\mu} - 1 + cB$ for some $0 < c < 1$, which in turn gives the estimates

$$A = O(e^{\frac{p}{2\mu}})\,, \quad \frac{B}{A} = O(e^{-\frac{p}{2\mu}} + e^{-\frac{q}{2\mu}})\,, \quad \text{as } \mu \to 0.$$

Now we can rewrite (5.8) and use Taylor expansion to obtain

$$\begin{aligned} e^{p/2\mu} - 1 &= A\left(e^{-\rho^2} - 2\rho\int_{-\infty}^{-\rho} e^{-\zeta^2}d\zeta\right)\Big|_{\rho=\frac{B}{\sqrt{4\pi A}}} \\ &= A\Big(1 - \sqrt{\pi}\rho + O(\rho^2)\Big)\Big|_{\rho=\frac{B}{\sqrt{4\pi A}}} = A\left(1 - \frac{1}{2}\frac{B}{A} + O\left((\frac{B}{A})^2\right)\right). \end{aligned}$$

We conclude that

$$B = 1 - e^{(p-q)/2\mu}\,, \quad A = e^{p/2\mu} - 1 + B\left[\frac{1}{2} + O(e^{-\frac{p}{2\mu}} + e^{-\frac{q}{2\mu}})\right] \tag{5.9}$$

and that $A = e^{p/2\mu} + O(B)$ as $\mu \to 0$.

Next we consider the behavior of $u_{p,q}$ as the viscosity $\mu \to 0$. We write this $u_{p,q}$ in the form

$$u_{p,q}(x,t) = \frac{\frac{x}{t} + \frac{B}{A}\sqrt{\frac{\mu}{\pi t_0}}}{1 + \sqrt{\frac{t}{t_0}}\frac{e^{\frac{x^2}{4\mu t}}}{A}\left(1 - B\frac{1}{\sqrt{\pi}}\int_{-\infty}^{x/\sqrt{4\mu t}} e^{-y^2}dy\right)}.$$

It is clear that

$$\frac{x}{t} + \frac{B}{A}\sqrt{\frac{\mu}{\pi t_0}} \to 0\,,$$

$$\frac{1}{A}e^{\frac{x^2}{4\mu t}}\left(1 - B\frac{1}{\sqrt{\pi}}\int_{-\infty}^{x/\sqrt{4\mu t}} e^{-y^2}dy\right) \sim \begin{cases} \frac{1}{A}e^{\frac{x^2}{4\mu t}}\,, & x < 0, \\ \frac{1-B}{A}e^{\frac{x^2}{4\mu t}}\,, & x > 0, \end{cases}$$
$$\sim \begin{cases} e^{\frac{1}{4\mu t}(x^2 - 2pt)}\,, & x < 0, \\ e^{\frac{1}{4\mu t}(x^2 - 2qt)}\,, & x > 0, \end{cases}$$

and, as a result,

$$u_{p,q}(x,t) \sim \begin{cases} 0, & x < -\sqrt{2pt}, \\ x/t, & -\sqrt{2pt} < x < \sqrt{2qt}, \\ 0, & x > \sqrt{2qt}, \end{cases} \tag{5.10}$$

as $\mu \to 0$. This validates the terminology diffusive N-wave for $u_{p,q}$.

The long-time behavior of $u_{p,q}(x,t)$ is also easily computed. A simple inspection shows that as $t \to \infty$

$$\varphi_{p,q}(x,t) \sim 1 - \frac{B}{\sqrt{4\pi\mu t}}\int_{-\infty}^{x} e^{-y^2/4\mu t}dy\,,$$

within leading order, and that $u_{p,q}$ has the structure of a diffusion wave of mass M.

If $M > 0$, the area of the negative part of the solution diminishes as $t \to \infty$. The time at which most of the negative area has almost disappeared can be estimated as follows: First, (5.7) is written in the form

$$e^{p(t)/2\mu} = 1 + A\sqrt{\frac{t_0}{t}}e^{-\frac{t}{t_0}\frac{B^2}{4\pi A^2}} - \frac{B}{\sqrt{\pi}}\int_{-\infty}^{-\sqrt{\frac{t}{t_0}}\frac{B}{\sqrt{4\pi}A}} e^{-\zeta^2}d\zeta. \tag{5.11}$$

We substitute $t = t_0A^2/B^2$ and obtain

$$e^{p(t)/2\mu} = 1 + |B|e^{-\frac{1}{4\pi}} - \frac{B}{\sqrt{\pi}}\int_{-\infty}^{-\frac{\mathrm{sign(B)}}{\sqrt{4\pi}}} e^{-\zeta^2}d\zeta. \tag{5.12}$$

If $q > p$, then $B \sim 1$, and we get

$$e^{p(t)/2\mu} \sim 1 + e^{-\frac{1}{4\pi}} - \frac{1}{\sqrt{\pi}}\int_{-\infty}^{-\frac{1}{\sqrt{4\pi}}} e^{-\zeta^2}d\zeta \sim 1.579$$

so that

$$p(t_0A^2/B^2) \sim 2\mu \ln(1.579) \sim 0.913\,\mu.$$

On the other hand, if $M < 0$ (equivalently, $p > q$), it is the area of the positive part of the solution that diminishes. The critical time is now estimated as follows: For $p > q$ and $\mu \ll 1$, we have $B \sim -e^{(p-q)/2\mu} < 0$. Dividing both sides of (5.12) by $|B|$, we obtain

$$e^{p(t)/2\mu}/|B| = 1/|B| + e^{-\frac{1}{4\pi}} + \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\frac{1}{\sqrt{4\pi}}} e^{-\zeta^2} d\zeta.$$

In turn,

$$e^{q(t)/2\mu} = e^{(p(t)-(p-q))/2\mu} \sim e^{-\frac{1}{4\pi}} + \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\frac{1}{\sqrt{4\pi}}} e^{-\zeta^2} d\zeta \sim 1.579,$$

and

$$q(t_0A^2/B^2) \sim 0.913\,\mu.$$

Remark 5.1. Because of the invariance of Burgers under translations $t \to t + a$, the functions $\varphi_{p,q}(x, t+1)$ and $u_{p,q}(x, t+1)$ are also special solutions associated to diffusive N-waves. The latter can be expressed (for $t_0 = 1$) in the form

$$\begin{aligned} \varphi_{p,q}(x, t+1) &= \psi_{p,q}\left(\frac{x}{\sqrt{4\mu(t+1)}}, t\right), \\ u_{p,q}(x, t+1) &= \sqrt{\frac{\mu}{t+1}}\, v_{p,q}\left(\frac{x}{\sqrt{4\mu(t+1)}}, t\right), \end{aligned} \tag{5.13}$$

where

$$\begin{aligned} \psi_{p,q}(\xi, t) &= 1 - \frac{B}{\sqrt{\pi}} \int_{-\infty}^{\xi} e^{-\zeta^2} d\zeta + A\frac{1}{\sqrt{t+1}} e^{-\xi^2}, \\ v_{p,q}(\xi, t) &= \frac{\frac{B}{\sqrt{\pi}} e^{-\xi^2} + 2A\frac{1}{\sqrt{t+1}} \xi e^{-\xi^2}}{1 - \frac{B}{\sqrt{\pi}} \int_{-\infty}^{\xi} e^{-\zeta^2} d\zeta + A\frac{1}{\sqrt{t+1}} e^{-\xi^2}}. \end{aligned} \tag{5.14}$$

The form (5.13) of the diffusive N-waves motivates an explicit solution for the Cauchy problem of the viscous Burgers equation, which is carried out in section 6. A survey of this solution and comparison with $v_{p,q}$ indicates that while the diffusion wave is the $t \to \infty$ asymptotic profile for the viscous Burgers, the diffusive N-wave $u_{p,q}$ gives a more accurate description of the behavior in the large time regime, which encompasses the very long-time behavior and the leading order correction and is valid for a substantially longer time interval.

6. An explicit solution of the viscous Burgers equation.

The objective of this section is to derive an explicit solution of the Cauchy problem for the viscous Burgers equation

$$\begin{aligned} &u_t + uu_x = \mu u_{xx}, \quad x \in \mathbb{R}, \quad \mu, t > 0, \\ &u(x, 0) = u_0(x), \qquad x \in \mathbb{R}. \end{aligned} \tag{6.1}$$

Our approach hinges on the invariance properties of the viscous Burgers equation, a self-similar variant of the Cole–Hopf transformation, and the exact solvability of a Fokker–Planck-type of equation. It yields an explicit solution of (6.1) in terms of Hermite polynomials.

Step 1. First, we apply to (6.1) the change of variables

$$u(x,t) = \sqrt{\frac{\mu}{t+1}}\, v\left(\frac{x}{\sqrt{4\mu(t+1)}}, t\right). \tag{6.2}$$

Then the function $v(\xi, t)$ of the similarity variable $\xi = \frac{x}{\sqrt{4\mu(t+1)}}$ satisfies the Cauchy problem

$$\begin{aligned} &(t+1)v_t + \left(-\tfrac{1}{2}\xi v + \tfrac{1}{4}v^2 \right)_\xi = \tfrac{1}{4}v_{\xi\xi}, \qquad && \xi \in \mathbb{R}, \quad t > 0, \\ &v(\xi, 0) = v_0(\xi), && \xi \in \mathbb{R}, \end{aligned} \tag{6.3}$$

with initial data

$$v_0(\xi) = \frac{1}{\sqrt{\mu}} u_0(\sqrt{4\mu}\xi). \tag{6.4}$$

The transformation is motivated by the invariance properties of the viscous Burgers equation and the form of the special solutions termed diffusive N-waves in section 5. Despite their similarity, the transformation (1.4) used in sections 2–4 differs in the dependence on viscosity and should not be confused with (6.2). In problem (6.3) the sole dependence on viscosity is through the initial data.

Step 2. We apply to (6.3) a variant of the Cole–Hopf transformation. Let

$$V(\xi, t) = \int_{-\infty}^{\xi} v(\zeta, t)\, d\zeta,$$

and introduce $\psi(\xi, t)$ so that

$$V = -\ln\psi, \qquad v = -\frac{\psi_\xi}{\psi}. \tag{6.5}$$

A calculation shows that V satisfies

$$(t+1)V_t - \frac{1}{2}\xi V_\xi + \frac{1}{4}V_\xi^2 = \frac{1}{4}V_{\xi\xi}, \tag{6.6}$$

and ψ satisfies the initial value problem

$$4(t+1)\psi_t = \psi_{\xi\xi} + 2\xi\psi_\xi, \qquad \xi \in \mathbb{R},\ t > 0, \tag{6.7}$$

with data

$$\begin{aligned} &\psi(\xi, 0) = \bar{\psi}(\xi) := e^{-V_0(\xi)}, \\ &\text{where} \quad V_0(\xi) = \int_{-\infty}^{\xi} v_0(\zeta) d\zeta = \frac{1}{2\mu}\int_{-\infty}^{\sqrt{4\mu}\xi} u_0(y)dy \\ &\qquad\qquad = \frac{M}{2\mu} - \frac{1}{2\mu}\int_{\sqrt{4\mu}\xi}^{\infty} u_0(y)dy. \end{aligned} \tag{6.8}$$

Step 3. Next we solve the initial value problem consisting of (6.7) with initial data

$$\psi(\xi,0) = \psi_0(\xi)\,, \qquad \text{with} \qquad e^{\frac{\xi^2}{2}}\psi_0 \in L^2(\mathbb{R}) \tag{6.9}$$

via separation of variables. This leads to the issue of finding the eigenvalues and eigenfunctions of the boundary value problem

$$g'' + 2\xi g' = \lambda g\,, \qquad -\infty < \xi < \infty. \tag{6.10}$$

The problem (6.10) turns out to have a discrete spectrum, associated with the Hermite polynomials.

The Hermite polynomials (see Szegö [19, Chapter V]) are the solutions $y = H_n(\xi)$, $n = 0, 1, 2, \dots$, of the boundary value problem

$$y'' - 2\xi y' + 2ny = 0\,, \qquad -\infty < \xi < \infty.$$

H_n are polynomials of degree n and are generated from the relation

$$H_n(\xi) = (-1)^n e^{\xi^2} \frac{d^n}{d\xi^n}\big(e^{-\xi^2}\big)\,.$$

The first few of them are $H_0 = 1$, $H_1 = 2\xi$, $H_2 = 4\xi^2 - 2$, and so on. They satisfy the orthogonality conditions

$$\int_{-\infty}^{\infty} H_m(\xi)\, H_n(\xi)\, e^{-\xi^2}\, d\xi = 2^n\, n!\, \sqrt{\pi}\, \delta_{nm}\,,$$

and the system $\{H_n(\xi)e^{-\frac{\xi^2}{2}}\}_{n=0}^{\infty}$ is a complete orthogonal system in $L^2(\mathbb{R})$.

Using these properties, it can be seen that the eigenvalues and eigenfunctions of (6.10) are $\lambda_n = -2(n+1)$ and $g_n(\xi) = e^{-\xi^2}H_n(\xi)$, $n = 0, 1, 2, \dots$. Moreover, the solution of (6.7)–(6.9) is given in the form of a series

$$\psi(\xi,t) = \sum_{n=0}^{\infty} a_n (t+1)^{-\frac{n+1}{2}} H_n(\xi) e^{-\xi^2}, \tag{6.11}$$

where a_n are determined by

$$\psi_0(\xi)\, e^{\frac{\xi^2}{2}} = \sum_{n=0}^{\infty} a_n H_n(\xi) e^{-\frac{\xi^2}{2}}\,.$$

For $\psi_0 e^{\frac{\xi^2}{2}} \in L^2(\mathbb{R})$ this problem is solvable, and the Fourier–Hermite coefficients are determined by the formula

$$a_n = \frac{1}{2^n\, n!\, \sqrt{\pi}} \int_{-\infty}^{\infty} \psi_0(\xi) H_n(\xi) d\xi\,. \tag{6.12}$$

Remark 6.1. The Hermite polynomials also appear in the eigenfunctions of the eigenvalue problem (sometimes called Hermite functions)

$$z'' + (2n+1-\xi^2)z = 0\,, \quad z_n(\xi) = H_n(\xi)e^{-\frac{\xi^2}{2}}\,, \quad n = 0, 1, 2, \dots\,,$$

which is associated with the problem of the harmonic oscillator in quantum mechanics. The eigenfunctions of the problem at hand are different from the ones above. The operator in (6.10) can be thought of as the integrated version of a Fokker–Planck-type operator.

Now consider the problem (6.7) with initial data

$$\psi(\xi,0)=\bar{\psi}(\xi)\,,\qquad \bar{\psi}(\xi)\to a \text{ as } \xi\to-\infty,\qquad \bar{\psi}(\xi)\to b \text{ as } \xi\to\infty. \tag{6.13}$$

The steady states $\psi^\infty(\xi)$ of (6.7)–(6.13) solve

$$\psi^\infty_{\xi\xi}+2\xi\psi^\infty_\xi=0\,,\qquad \psi^\infty(-\infty)=a\,,\ \ \psi^\infty(\infty)=b\,,$$

and are given by

$$\psi^\infty(\xi)=a+\frac{b-a}{\sqrt{\pi}}\int_{-\infty}^{\xi}e^{-\zeta^2}d\zeta\,. \tag{6.14}$$

By superposition, it is possible to solve (6.7) with initial data

$$\bar{\psi}(\xi)=\psi^\infty(\xi)+\psi_0(\xi)\,,\qquad \text{where}\quad \psi_0 e^{\frac{\xi^2}{2}}\in L^2(\mathbb{R}).$$

Its solution is given in the form

$$\psi(\xi,t)=\psi^\infty(\xi)+\sum_{n=0}^{\infty}a_n(t+1)^{-\frac{n+1}{2}}H_n(\xi)e^{-\xi^2}, \tag{6.15}$$

where the Fourier coefficients a_n are computed from (6.12).

Step 4. Returning now to Burgers, we assume that $u_0(x)=O(e^{-x^2})$ as $|x|\to\infty$. Then $v_0(\xi)=O(e^{-\xi^2})$ as $|\xi|\to\infty$. (The orders will, in general, depend on the viscosity.) The initial data $\bar{\psi}$ are given in (6.8) and satisfy $\bar{\psi}(-\infty)=1$ and $\bar{\psi}(\infty)=e^{-\frac{M}{2\mu}}$. We define the associated diffusion wave

$$\begin{aligned}\psi^\infty_M&=1-\frac{1-e^{-\frac{M}{2\mu}}}{\sqrt{\pi}}\int_{-\infty}^{\xi}e^{-\zeta^2}d\zeta\\ &=e^{-\frac{M}{2\mu}}+\frac{1-e^{-\frac{M}{2\mu}}}{\sqrt{\pi}}\int_{\xi}^{\infty}e^{-\zeta^2}d\zeta.\end{aligned} \tag{6.16}$$

Consider

$$\psi_0=\bar{\psi}-\psi^\infty_M=e^{-V_0}-\psi^\infty_M,$$

and note that ψ_0 can be expressed as

$$\begin{aligned}\psi_0(\xi)&=e^{-\int_{-\infty}^{\xi}v_0d\zeta}-1+\frac{1-e^{-\frac{M}{2\mu}}}{\sqrt{\pi}}\int_{-\infty}^{\xi}e^{-\zeta^2}d\zeta\\ &=e^{-\frac{M}{2\mu}+\int_{\xi}^{\infty}v_0d\zeta}-e^{-\frac{M}{2\mu}}-\frac{1-e^{-\frac{M}{2\mu}}}{\sqrt{\pi}}\int_{\xi}^{\infty}e^{-\zeta^2}d\zeta.\end{aligned}$$

Using the inequalities $|e^x-1|\le 2x$ for $|x|\ll 1$ and the decay $v_0(\xi)=O(e^{-\xi^2})$ as $|\xi|\to\infty$, we see that $\psi_0(\xi)=O(e^{-\xi^2})$ as $|\xi|\to\infty$. We apply the results of the previous section and see that ψ is given by (6.15). As a result, the solution of (6.1) is given by the formula

$$u(x,t) = \sqrt{\frac{\mu}{t+1}}\, v\left(\frac{x}{\sqrt{4\mu(t+1)}}, t\right),$$

$$\text{where}\quad v(\xi,t) = -\frac{\partial_\xi \psi_M^\infty + \displaystyle\sum_{n=0}^{\infty} a_n (t+1)^{-\frac{n+1}{2}} \partial_\xi \left(H_n(\xi) e^{-\xi^2}\right)}{\psi_M^\infty + \displaystyle\sum_{n=0}^{\infty} a_n (t+1)^{-\frac{n+1}{2}} H_n(\xi) e^{-\xi^2}}. \tag{6.17}$$

ψ_M^∞ is given in (6.16), and the coefficients a_n are computed by

$$a_n = \frac{1}{2^n\, n!\, \sqrt{\pi}} \int_{-\infty}^{\infty} \left(e^{-V_0} - \psi_M^\infty\right) H_n d\xi\,. \tag{6.18}$$

In view of the form of the Hermite polynomials, the coefficients a_n may also be expressed in terms of moments of the function $\psi_0 = e^{-V_0} - \psi_M^\infty$.

An inspection of (6.17) shows that as $t \to \infty$

$$v \sim -\frac{\partial_\xi \psi_M^\infty}{\psi_M^\infty},$$

which is the asymptotic profile of a diffusion wave of mass M. The next order approximation is (recall that $H_0 = 1$)

$$v(\xi,t) \sim -\frac{\partial_\xi \psi_M^\infty - 2a_0 (t+1)^{-\frac{1}{2}} \xi e^{-\xi^2}}{\psi_M^\infty + a_0 (t+1)^{-\frac{1}{2}} e^{-\xi^2}},$$

which is that of a diffusive N-wave (compare with (5.14)). The coefficient a_0 is computed by the formula

$$a_0 = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-V_0} - \psi_M^\infty \, d\xi\,.$$

REFERENCES

[1] S. Angenent, *The zero set of a solution of a parabolic equation*, J. Reine Angew. Math., 390 (1988), pp. 79–96.

[2] C. M. Dafermos, *Generalized characteristics and the structure of solutions of hyperbolic conservation laws*, Indiana Univ. Math. J., 26 (1977), pp. 1097–1119.

[3] C. M. Dafermos, *Asymptotic behavior of solutions of evolution equations*, in Nonlinear Evolution Equations, M. G. Crandall, ed., Academic Press, New York, 1978, pp. 103–124.

[4] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, New York, 2000.

[5] M. Escobedo and E. Zuazua, *Large time behavior for convection-diffusion equations in R^N*, J. Funct. Anal., 100 (1991), pp. 119–161.

[6] M. Escobedo, J. L. Vazquez, and E. Zuazua, *Asymptotic behaviour and source-type solutions for a diffusion-convection equation*, Arch. Ration. Mech. Anal., 124 (1993), pp. 43–65.

[7] E. Hopf, *The partial differential equation $u_t + uu_x = \mu u_{xx}$*, Comm. Pure Appl. Math., 3 (1950), pp. 201–230.

[8] S. N. Kruzhkov, *First order quasilinear equations in several independent variables*, Math. USSR Sb. 10 (1970), pp. 217–243.

[9] S. N. Kruzhkov, *Quasilineare Gleichungen erster Ordnung mit mehreren unabhaengigen Veraenderlichen*, Mat. Sb., 123 (1970), pp. 228–255.

[10] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Lectures in Mathematics ETH Zürich, Birkhäuser-Verlag, Basel, 1990.

[11] G. Lieberman, *Second Order Parabolic Differential Equations*, World Scientific, River Edge, NJ, 1996.

[12] T.-P. Liu, *Invariants and asymptotic behavior of solutions of a conservation law*, Proc. Amer. Math. Soc., 71 (1978), pp. 227–231.

[13] T.-P. Liu and M. Pierre, *Source-solutions and asymptotic behavior in conservation laws*, J. Differential Equations, 51 (1984), pp. 419–441.

[14] T.-P. Liu, A. Matsumura, and K. Nishihara, *Behaviors of solutions for the Burgers equation with boundary corresponding to rarefaction waves*, SIAM J. Math. Anal., 29 (1998), pp. 293–308.

[15] H. Matano, *Nonincrease of the lap number of a solution for a one-dimensional semi-linear parabolic equation*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 29 (1982), pp. 401–441.

[16] O. Oleinik, *Discontinuous solutions of nonlinear differential equations*, Uspekhi Mat. Nauk (N.S.), 12 (1957), pp. 3–73 (in Russian).

[17] O. Oleinik, *Discontinuous solutions of nonlinear differential equations*, Amer. Math. Soc. Transl. (2), 26 (1963), pp. 95–172.

[18] D. H. Sattinger, *On the total variation of solutions of parabolic equations*, Math. Ann., 183 (1969), pp. 78–92.

[19] G. Szegö, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ. 23, AMS, Providence, RI, 1939.

[20] L. Tartar, *Une introduction a la théorie matheématique des systèmes hyperboliques de lois de conservation*, Pubblicazioni CNR Instituto di Analisi Numerica, 682, Pavia, Italy, 1989.

[21] G. Whitham, *Linear and Nonlinear Waves*, Pure Appl. Math., Wiley Interscience, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 33, No. 3, pp. 634–648

ARE NATURAL IMAGES OF BOUNDED VARIATION?*

YANN GOUSSEAU† AND JEAN-MICHEL MOREL†

Abstract. The bounded variation assumption is the starting point of many methods in image analysis and processing. However, one common drawback of these methods is their inability to handle textures and small structures properly. Here we precisely show why natural images are incompletely represented by BV functions. Through an experimental study of the distribution of bilevels of natural images, we show that their total variation blows up to infinity with the increasing of resolution. To reach these conclusions, we compute bounds on the total variation, and we model convolution and sampling under quite general assumptions.

Key words. natural images, bounded variation, total variation, bilevels, power laws, size distribution, wavelets

AMS subject classifications. 26B15, 26B30, 28A75, 28A80, 65C20, 68U10

PII. S0036141000371150

1. Introduction.

This paper addresses the question of whether natural images may be represented as functions of bounded variation. The question is of relevance because of the wide use of the space BV in image modeling. Roughly speaking, the space BV is the space of functions whose weak derivative is a measure with finite total variation and is a straightforward space for images since it contains characteristic functions of simple sets, thus enabling the representation of edges. The BV assumption is the starting point of different approaches in image restoration [Rud87], [ROF92], [CW98], [Mal99a], image segmentation [Amb89], image deconvolution [RO94], [KMR99], optical flow computation [ADK99], or image compression. These methods have proven very efficient, especially in dealing with one-dimensional discontinuity in images. However, one common drawback is their inability to handle textures properly. In particular, restoration or deconvolution in BV leads to the smoothing out of textures, and segmentation procedures in BV fail to isolate textured areas. A recent paper [Nik00] yields mathematical proofs of the stair-casing effect, according to which BV minimization tends to create constant patches in images, thus eliminating textural effects.

In this paper, we show that this phenomenon can be explained by the fact that natural images are not of bounded variation. Our approach combines an experimental program we performed on the distribution of homogeneous and connected regions in images, the sections, and a theoretical result bounding from below the BV norm of two-dimensional functions according to the distribution of their sections. We formalize the link between experiments on discrete images and the mathematical results by using a simple model of convolution and sampling for the formation of numerical images. We point out here that another method to estimate the BV norm of images relies on the study of wavelet coefficients. In order to study locally the (ir)regularity of a signal, one may investigate the decay of wavelet coefficients at a certain location (see [Mey93], [Mal99b] for an introduction to wavelet decompositions). More recently,

*Received by the editors April 20, 2000; accepted for publication (in revised form) February 28, 2001; published electronically October 4, 2001.
http://www.siam.org/journals/sima/33-3/37115.html

†CMLA, ENS Cachan, 61 av. du Président Wilson, 94235 Cachan Cedex, France (gousseau@cmla.ens-cachan.fr).

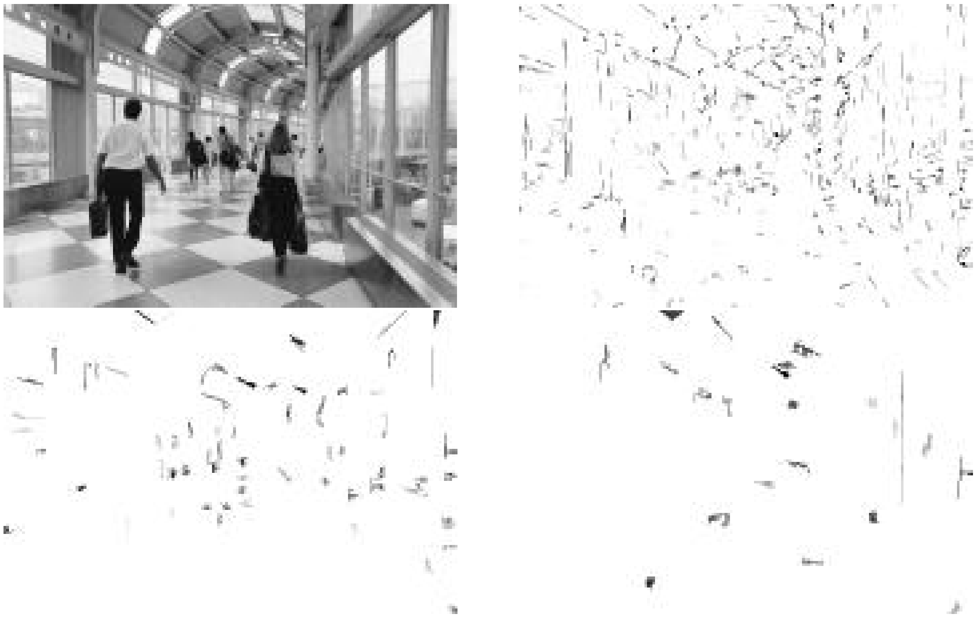


Fig. 1. *Airport,* 510 × 343 *image, with some of its sections for* $k = 10$. *Top right sections are of size between* 10 *and* 20 *pixels; bottom left between* 40 *and* 50 *pixels; bottom right between* 80 *and* 90 *pixels. Different gray levels correspond to sections from different* k*-bilevels.*

the link between the global decay of wavelet coefficients and the BV norm has been studied so that, in some cases, this decay permits us to decide whether or not a function belongs to the space BV; see [CDPX99], [Oru98]. We shall compare our method and the wavelet method to decide whether an image is in BV or not. As we shall see, the geometric measurements we perform seem more accurate and permit us to show that natural images are not of bounded variation.

The paper is organized as follows: in section 2, we recall our results about the power law distribution of sections' area and perimeter. In section 3, we recall basic definitions and properties related to the space BV of functions with bounded variation. In section 4, we establish a link between the distribution of section size and the BV norm of functions of $\mathbb{R}^2$; in section 5, we show that by combining the experimental results of section 2 and the theoretical results of section 4, we can conclude that natural images are not of bounded variation. Eventually, in section 6, we compare our conclusions with recent results on the decay of wavelet coefficients.

2. The distribution of bilevels in natural images. In previous papers (see [AGM99], [Gou00]), we explored the statistics of homogeneous and connected regions of natural images. The most remarkable fact is that the distributions of their area and perimeter are very well approximated by a power law. More precisely, we consider a digital image I, whose gray levels are between 0 and N. For an integer k, we call a k-bilevel of I any of the binary images defined by

$$I_l(i,j) = \begin{cases} 1 & \text{if } I(i,j) \in \left[(l-1)\frac{N}{k}, l\frac{N}{k}\right), \\ 0 & \text{otherwise} \end{cases}$$

for l varying from 1 to k. We then define a section to be a connected component of a set $\{(i,j)/I_l(i,j) = 1\}$ for some l. For each integer a, we define $f(a)$ to be the number

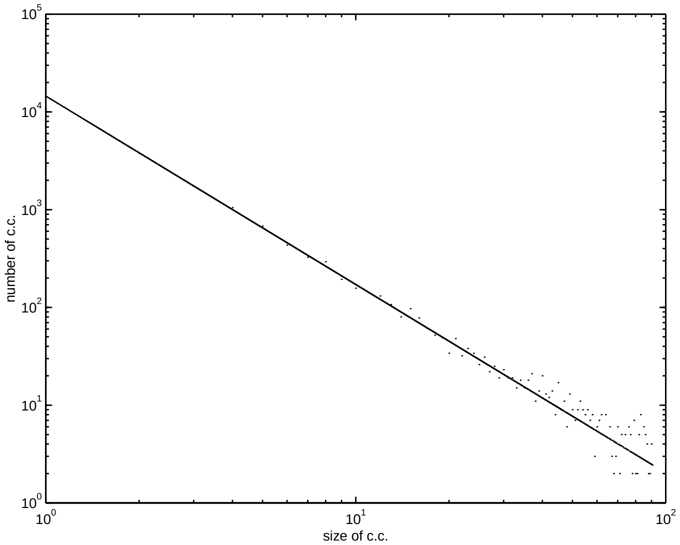


FIG. 2. *Function f (section area distribution) for the airport image, Figure* 1, *in* log-log *coordinates.*

TABLE 1

Average results for the distribution of area of sections on 100 *images from the van Hateren database. We denote by $\langle\alpha\rangle$ and $\langle E\rangle$ the mean values of α and E respectively, std α is the standard deviation of α;* min*(α) and* max*(α) are the minimum and maximum values for α.*

k	$\langle\alpha\rangle$	std α	max α	min α	$\langle E\rangle$	max E
16	1.85	0.19	2.20	1.39	0.37	0.49
14	1.83	0.19	2.18	1.36	0.37	0.54
12	1.83	0.19	2.15	1.37	0.37	0.54
10	1.81	0.18	2.12	1.27	0.37	0.50
8	1.80	0.17	2.26	1.32	0.38	0.59

of sections with area a (in pixels). Our experiments show that

$$f(a) \approx \frac{C}{a^{\alpha}}, \tag{2.1}$$

where C and α are image dependent constants. Moreover, in most images, α is close to 2.

We will not recall here all the experimental results, and we refer to the previously mentioned papers for more details, but we will give some examples. For each image, we fit a straight line with slope α to the function f in log-log coordinate, minimizing the least squares distance. We also compute the least squares error E. In Figure 1, we display a digital image and some of its sections, and in Figure 2 we display the corresponding fit for f. Similar graphics are obtained for all considered digital images (either from a digital camera, scanned images, or calibrated images). In Table 1, we display results averaged over 100 calibrated images from a database collected by van Hateren, freely available at http://hlab.phys.rug.nl/imlib/; see [vHvdS98]. The most noticeable fact here is the proximity of α to 2, and the fact that for all images and some value of k, α is larger than 1.5, a fact that will be proven relevant in what follows for estimating the BV norm of images.

Table 2

Average results for the distribution of perimeter of sections on 100 *images from the van Hateren database. Notations are the same as for Table* 1.

k	$\langle\beta\rangle$	std β	max β	min β	$\langle E\rangle$	max E
16	2.35	0.28	2.57	2.04	0.36	0.42
14	2.42	0.29	2.60	2.10	0.29	0.35
12	2.38	0.33	2.63	1.99	0.42	0.51
10	2.46	0.15	2.62	2.10	0.31	0.39
8	2.36	0.14	2.49	2.04	0.37	0.41

In [AGM99], [Gou00], we also studied the distribution of the perimeters of sections, which are also distributed according to a power law with an exponent β usually between 2 and 3. In Table 2, we display results similar to those in Table 1 for the values of β (still computed by minimizing the least squares error in log-log coordinates) on the previously mentioned images database.

Are small sections due to noise or microtextures? In Figure 1, we have shown some of the small sections from which the size statistics are estimated. We do that for the following purpose: it might be objected to the observed size laws that their small scale behavior is due to the caption device and not to the underlying "natural" image. Thus it is very relevant to look at the sections and decide whether they are due to digitization noise, to some microtexture, or to the inherent geometric structure of the image. In Figure 1, we can check that most small sections arise on contrasted parts of the image (the so called "edges") and that their shape coincides with those edges. We can also rule out a Gibbs phenomenon: it multiplies the edge contribution to the bounded variation norm by a fixed constant factor. We have shown only one example, but we have chosen a kind of example for which the BV model should be very likely, since the whole scene is a geometric human-made scene with as little texture as possible. All other images we have checked confirm this interpretation: small sections correspond to objects or pieces of objects, or pieces of contours. They are not at all uniformly distributed over the image, as would happen with noise.

3. Functions of bounded variation and sets of finite perimeter.

In this section, we recall some basic facts about functions of bounded variation. Let I be a bounded function defined on a domain (e.g., rectangular) $\Omega \subset \mathbb{R}^2$. I is in $BV(\Omega)$, the space of functions with bounded variations, if

$$||I||_{BV} \overset{\text{def}}{=} \int_\Omega |DI| < +\infty,$$

where the gradient DI is to be understood in a weak sense (see [Zie89]):

$$\int_\Omega |DI| = \sup\left\{\int_\Omega I \text{div}\phi \mid \phi \in C_c^1(\Omega), |\phi| < 1\right\},$$

where $C_c^1(\Omega)$ is the space of continuously differentiable functions with compact support, defined from Ω to $\mathbb{R}^2$. Actually the usual BV norm is defined as the sum of $\int |DI|$, the total variation, and the L^1 norm, $\int |I|$. We consider only the total variation (which is not a norm), and write it $||.||_{BV}$.

We will be interested in a more geometric characterization of $BV(\Omega)$. For $\lambda \in \mathbb{R}$, define the level set of I with level λ by

$$\chi_\lambda I = \{x, I(x) \geq \lambda\}.$$

Now (see [EG92]) recall that a set $E \in \Omega$ is of finite perimeter per(E) if

$$\mathrm{per}(E) \overset{\mathrm{def}}{=} ||\mathbb{1}_E||_{BV} \leq +\infty,$$

where $\mathbb{1}_E$ is the characteristic function of the set E. This definition generalizes the usual definition of the boundary length, in the sense that both definitions are equivalent in the case of a set with piecewise regular boundary. If a function has bounded variation, then, for almost every $\lambda \in \mathbb{R}$, $\chi_\lambda I$ is a set with finite perimeter and (coarea formula (see [EG92])),

$$||I||_{BV} = \int_{\mathbb{R}} \mathrm{per}(\chi_\lambda I) d\lambda. \tag{3.1}$$

Conversely, if per$(\chi_\lambda I)$ is finite (for almost all λ) and the preceding integral is finite, then I has bounded variation. We also recall that, by the classical isoperimetric inequality, we have for every set O with finite perimeter,

$$\mathrm{per}(O) \geq 2\pi^{\frac{1}{2}} \nu(O)^{\frac{1}{2}}, \tag{3.2}$$

where $\nu(O)$ denotes the Lebesgue measure of O.

The BV space is a very straightforward space for images. First, if images are neither continuous nor strictly differentiable, it seems reasonable to assume them to be in a space where they are weakly differentiable. Moreover, the occlusion phenomenon is responsible for one-dimensional discontinuities which prevent the weak derivatives of images from being integrable, thus forcing images out of any Sobolev space. Such a simple image as a white disk on a black background belongs to the space BV, which is the natural space to perform calculus of variations on functions whose one-dimensional discontinuities have finite length (see [Amb89]). Now, a first way for an image not to be in this space is to have level lines with infinite length. For instance, the characteristic functions of two-dimensional sets with fractal boundaries will not be of bounded variation. There is also another way for a function not to be in BV. Each of its level lines may be of finite perimeter, while the sum of these level lines' perimeters is infinite. As we will see in the next two sections, this is what happens for natural images, in which, in a precise sense, small objects are too numerous for the function to be of bounded variation.

4. A lower bound for the *BV* norm.

In the following, we shall consider sections of the image. We always assume that the image I satisfies $0 \leq I(x) \leq C$. We first fix two parameters γ, λ, with $0 \leq \lambda \leq \gamma$. For any $n \in \mathbb{N}$, we consider the bilevel sets of I

$$\{x, \lambda + (n-1)\gamma \leq I(x) < \lambda + n\gamma\} = \chi_{\lambda+(n-1)\gamma} I \setminus \chi_{\lambda+n\gamma} I.$$

We call the (γ, λ)-section of I any set which is a connected component of a bilevel set $\chi_{\lambda+(n-1)\gamma} I \setminus \chi_{\lambda+n\gamma} I$ for some n. We denote each one of the components by $S_{\gamma,\lambda,i}$ for $i \in J(\gamma, \lambda)$, a set of indices. Notice that the (γ, λ)-sections are disjoint, and their union is the image domain Ω,

$$\bigcup_{i \in J(\gamma,\lambda)} S_{\gamma,\lambda,i} = \Omega. \tag{4.1}$$

There are several ways to define the connected components of a set with finite perimeter, since such a set is defined up to a set with zero Lebesgue measure. One can prove [ACMM99] that a definition of connected components for a set with finite perimeter permits the following statements. (Recall that ν is the two-dimensional Lebesgue measure and per is the perimeter.)

Definition 4.1. *Let X be a set with finite perimeter in $\mathbb{R}^2$. We say that X is not decomposable if we cannot write it as $X = Y \cup Z$ with $\nu(Y) > 0$, $\nu(Z) > 0$, $\nu(X) = \nu(Y) + \nu(Z)$, and $per(X) = per(Y) + per(Z)$.*

Theorem 4.2. *Each set of finite perimeter X admits a unique decomposition*

$$X = \cup_n X_n,$$

where the union is finite or countable, and such that

(i) *each X_n is not decomposable,*
(ii) *for each n, $\nu(X_n) > 0$,*
(iii) *$per(X) = \sum_n per(X_n)$.*

This definition matches the usual requirements of connectivity, in particular, if for $x \in X$, cc(x,X) is the component relative to X that contains x, $X \subset Y$ implies $\mathrm{cc}(x, X) \subset \mathrm{cc}(x, Y)$.

We need this definition because (iii) will enable us to use the distribution of sections, as experimentally observed in section 2, to bound the BV norm of images from below. We denote by $J(n) \subset J(\gamma, \lambda)$ the set of indices of sections which are connected components of $\chi_{\lambda+(n-1)\gamma} I \setminus \chi_{\lambda+n\gamma} I$. Note that by classical results on BV functions, for each γ, $\chi_{\lambda+(n-1)\gamma} I \setminus \chi_{\lambda+n\gamma} I$ has finite perimeter for almost every λ. As an obvious consequence of Proposition 4.2, we have the following corollary.

Corollary 4.3. *Let I belong to BV. Then for almost every λ,*

$$per(\chi_{\lambda+(n-1)\gamma} I \setminus \chi_{\lambda+n\gamma} I) = \sum_{i \in J(n)} per(S_{\lambda,\gamma,i}).$$

In order to estimate the BV norm of I, we shall need the following lemma.

Lemma 4.4. *If $B \subset A$ are two sets with finite perimeter, then*

$$per(A \setminus B) \leq per(A) + per(B).$$

Proof. Recall that $\mathrm{per}(A) = ||\mathbb{1}_A||_{BV}$. Then by the subadditivity of the BV norm, we deduce from

$$\mathbb{1}_{A \setminus B} = \mathbb{1}_A - \mathbb{1}_B$$

that

$$\mathrm{per}(A \setminus B) \leq \mathrm{per}(A) + \mathrm{per}(B). \qquad \square$$

In the following theorem, we analyze the statistics of sizes of sections as follows. We fix γ, that is, the overall contrast of considered sections, and for each $0 \leq \lambda \leq \gamma$, we count all sections $S_{\gamma,\lambda,i}$ which have an area between s_1 and s_2 with $0 < s_1 < s_2$. That is, we consider the integer

$$\mathrm{Card}\{i, s_1 \leq \nu(S_{\gamma,\lambda,i}) < s_2\}.$$

Note that this number is bounded since Ω is bounded and the sections are disjoint. We average this number over all λ's in $[0,\gamma]$ to obtain the function

$$f(\gamma, s_1, s_2) = \int_0^\gamma \mathrm{Card}\{i, s_1 \le |S_{\gamma,\lambda,i}| < s_2\} d\lambda.$$

Remark. To be able to define f, we made the assumption that

$$\mathrm{Card}\{i, s_1 \le \nu(S_{\gamma,\lambda,i}) < s_2\} \text{ is a measurable function of } \lambda. \tag{4.2}$$

We will suppose that for some $\gamma > 0$, this average number has a density $f(\gamma, s)$ with respect to s. That is,

$$\forall s > 0 \qquad \lim_{s_1 \uparrow s, s_2 \downarrow s} \frac{f(\gamma, s_1, s_2)}{s_1 - s_2} = f(\gamma, s). \tag{4.3}$$

Then we have the following bound for the BV norm of I.

THEOREM 4.5. *Let I be in $BV(\Omega)$. Assume that there exists some $\gamma > 0$ such that (4.2) and (4.3) hold (i.e., the average number of sections with area s, for $0 \le \lambda \le \gamma$, has a density $f(\gamma, s)$); then*

$$||I||_{BV} \ge \pi^{\frac{1}{2}} \int_0^{\nu(\Omega)} s^{\frac{1}{2}} f(\gamma, s) ds. \tag{4.4}$$

Proof. Applying Corollary 4.3 and Lemma 4.4,

$$\begin{aligned}
||I||_{BV} &= \int_{\mathbb{R}} \mathrm{per}\{x, I(x) \ge \lambda\} d\lambda \\
&= \frac{1}{2}\left(\int_{\mathbb{R}} \mathrm{per}\{x, I(x) \ge \lambda\} d\lambda + \int_{\mathbb{R}} \mathrm{per}\{x, I(x) \ge \lambda - \gamma\} d\lambda\right) \\
&\ge \frac{1}{2} \int_{\mathbb{R}} \mathrm{per}(\chi_{\lambda-\gamma} I \setminus \chi_\lambda I) d\lambda \\
&= \frac{1}{2} \sum_{n \in \mathbb{Z}} \int_{n\gamma}^{(n+1)\gamma} \mathrm{per}(\chi_{\lambda-\gamma} I \setminus \chi_\lambda I) d\lambda \\
&= \frac{1}{2} \int_0^\gamma \sum_{n \in \mathbb{Z}} \mathrm{per}(\chi_{\lambda+(n-1)\gamma} I \setminus \chi_{\lambda+n\gamma} I) d\lambda \\
&= \frac{1}{2} \int_0^\gamma \sum_{i \in J(\gamma,\lambda)} \mathrm{per}(S_{\gamma,\lambda,i}) d\lambda.
\end{aligned}$$

By the isoperimetric inequality (3.2), we therefore obtain

$$||I||_{BV} \ge \pi^{\frac{1}{2}} \int_0^\gamma \sum_{i \in J(\gamma,\lambda)} \nu(S_{\gamma,\lambda,i})^{\frac{1}{2}} d\lambda.$$

Then, for any $n \in \mathbb{N}^*$

$$||I||_{BV} \ge \pi^{\frac{1}{2}} \sum_{k=1}^{n-1} \left(\frac{\nu(\Omega)}{n} k\right)^{\frac{1}{2}} f\left(\gamma, \frac{\nu(\Omega)}{n} k, \frac{\nu(\Omega)}{n}(k+1)\right). \tag{4.5}$$

We introduce the functions

$$f_n = \sum_{k=1}^{n-1} \left(\frac{\nu(\Omega)}{n}k\right)^{\frac{1}{2}} f\left(\gamma, \frac{\nu(\Omega)}{n}k, \frac{\nu(\Omega)}{n}(k+1)\right) \frac{n}{\nu(\Omega)} \mathbb{1}_{\left[\frac{\nu(\Omega)}{n}k, \frac{\nu(\Omega)}{n}(k+1)\right)}.$$

We have

$$||I||_{BV} \geq \pi^{\frac{1}{2}} \int_0^{\nu(\Omega)} f_n(s)ds$$

and

$$\forall s_0 > 0 \qquad f_n(s_0) \underset{n\to+\infty}{\longrightarrow} (s_0)^{\frac{1}{2}} \gamma f(\gamma, s_0),$$

thanks to hypothesis (4.3). Therefore, by Fatou's lemma,

$$||I||_{BV} \geq \pi^{\frac{1}{2}} \int_0^{\nu(\Omega)} s^{\frac{1}{2}} f(\gamma, s)ds. \qquad \square$$

We can repeat the preceding analysis by assuming now that

$$g(\gamma, p_1, p_2) = \int_0^{\gamma} \text{Card}\{i, p_1 \leq \text{per}(S_{\gamma,\lambda,i}) \leq p_2\} d\lambda$$

has an average density $g(\gamma, p)$ with respect to p (once more assuming the cardinal we integrate is measurable). That is,

$$(4.6) \qquad \lim_{p_1\uparrow p, p_2\downarrow p} \frac{g(\gamma, p_1, p_2)}{p_2 - p_1} = g(\gamma, p).$$

Then we have the analogue of Theorem 4.5 for the perimeters of sections.

THEOREM 4.6. *Let I be in $BV(\Omega)$. Assume that there exists some $\gamma > 0$ such that* (4.6) *holds, i.e., the average number of sections with perimeter s, for $0 \leq \lambda \leq \gamma$, has a density $g(\gamma, p)$. Then*

$$(4.7) \qquad ||I||_{BV} \geq \frac{1}{2} \int_0^{+\infty} pg(\gamma, p)dp.$$

Proof. In the same way as before (without using the isoperimetric inequality) and fixing some $p_m > 0$,

$$\begin{aligned} ||I||_{BV} &\geq \frac{1}{2} \int_0^{\gamma} \sum_{i\in J(\gamma,\lambda)} \text{per}(S_{\gamma,\lambda,i}) d\lambda \\ &\geq \frac{1}{2} \sum_{k=1}^{n-1} \left(\frac{p_m}{n}k\right) g\left(\gamma, \frac{p_m}{n}k, \frac{p_m}{n}(k+1)\right). \end{aligned}$$

As before, this implies

$$||I||_{BV} \geq \frac{1}{2} \int_0^{+\infty} pg(\gamma, p)dp. \qquad \square$$

5. Application: Natural images are not of bounded variation.

5.1. The continuous framework. In this section, we draw the consequences of Theorems 4.5 and 4.6 for the images analyzed in section 2 by assuming that the observed distribution of sections approximates the distribution in continuous images. According to our experimental results, we suppose that the considered images satisfy

$$f(\gamma, s) = \frac{C}{s^{\alpha}}, \tag{5.1}$$

$$g(\gamma, p) = \frac{C}{p^{\beta}} \tag{5.2}$$

for some constants $\alpha > 0$, $\beta > 0$, where f is the density for the area distribution of the sections, and g is the density for the perimeters. This law has been experimentally checked for several values of $\gamma = \frac{256}{k}$ (the grey level width of the sections) and k ranging from 8 to 20; see section 2 and [Gou00]. We also checked that the value of α was almost unchanged when the bilevels were not defined from gray level 0, but from some gray level less than $\frac{256}{k}$ (that is, in the continuous model for different values of λ), and that when averaging the experimental density function over integer values of λ between 0 and $\gamma = \frac{256}{k}$, f and g still are power laws with the same exponent. Thus hypotheses (5.1) and (5.2) are valid. We emphasized here that the shapes and locations of small sections indicate that these are not due to noise, but to small structures and objects clearly present in the image, particularly, pieces of edges; see Figure 1. Moreover, Gaussian white noise leads to quite different statistics; see [Gou00].

Then, by Theorem 4.5, we have

$$||I||_{BV} \geq c \int_0^{\nu(\Omega)} \frac{Cs^{\frac{1}{2}}}{s^{\alpha}} ds$$

and, in the same way,

$$||I||_{BV} \geq c \int_0^{+\infty} \frac{Cp}{p^{\beta}} dp.$$

Thus, if we admit that (5.1) and (5.2) indeed hold for natural images when $s \to 0$, as is indicated by the experiments recalled in section 2, we obtain that the considered images are not in BV if $\alpha > \frac{3}{2}$ or $\beta > 2$, since the corresponding integrals are not finite. These values of α and β have been checked for all images of the database studied in section 2, for some value of k, except for some (3 out of 100) blurred images. The assumption that formulae (5.1) and (5.2) hold for small scales, below the scale of pixelization, is a strong one, but the experiment of section 2 indicates that the distribution is the same at all scales. Moreover, the next section will analyze the effect of pixelisation on the BV norm. Of course, there exists a cut-off scale in images, but the lower bounds we found for the BV norm indicate that the contribution of small scales to the value of this norm is unexpectedly large compared to the contribution of larger scales. As mentioned in the introduction, this should be related to the problem of the erasing of textures by variational methods minimizing the total variation of images. Indeed, the BV norm gives a large weight to small-scale textures that are known to disappear in the process of restoration (deblurring or denoising, for instance) by such variational methods; see [ROF92], [Mal99a].

5.2. Convolution and sampling. In this section, we draw the same conclusions as in the previous one about the BV norm of natural images, with a more rigorous interpretation of our experimental results. We give a practical application of the method of Theorem 4.5 in a case where its hypotheses are satisfied. We assume that discrete images, on which we performed the analysis of section 2, are obtained from continuous ones through convolution and sampling. We first show that when we compute the BV norm of a function after convolution with a rescaled smoothing kernel (under some regularity conditions) and sampling, we underestimate the actual value of the BV norm of the initial function. We write G for a two-dimensional function and C for the square whose lower left corner is (0,0) and whose upper right corner (1,1); n is an integer, and for a real number x, $[x]_n = [nx]/n$, $[nx]$ is the integer part of nx.

LEMMA 5.1. *Let I be a function in $BV(C)$, let $G_n(x,y) = G(nx, ny)$, define*

$$I_c = I * G_n,$$

and for $(x,y) \in C$

$$I_n(x,y) = I_c\left([x]_n + \frac{1}{2n}, [y]_n + \frac{1}{2n}\right).$$

Assume that there exist some functions a and b such that

$$\forall x, y \in C \qquad |G(x,y)| \leq a(x)b(y), \tag{5.3}$$

and assume there exists a constant K such that

$$\sup_{x,y} \sum_{i,j} a(x+i)b(y+j) \leq K. \tag{5.4}$$

Then

$$||I_n||_{BV(C)} \leq \sqrt{2}K||I||_{BV(C)}.$$

Proof. For i, j integer in $[0,n)$, define

$$I_{i,j} = I_c\left(\frac{i}{n} + \frac{1}{2n}, \frac{j}{n} + \frac{1}{2n}\right).$$

We have

$$||I_n||_{BV} = \sum_{i,j} \left(|I_{i,j} - I_{i-1,j}| + |I_{i,j} - I_{i,j-1}|\right).$$

Now, for all i, j,

$$\begin{aligned}
|I_{i,j} - I_{i-1,j}| &= \left|\int \left(I\left(\frac{i+1}{n} - x, \frac{j}{n} - y\right) - I\left(\frac{i}{n} - x, \frac{j}{n} - y\right)\right) G(nx, ny)\right| \\
&= \left|\int \int_0^1 \frac{\partial}{\partial x} I\left(\frac{i+t}{n} - x, \frac{j}{n} - y\right) G(nx, ny)\right| \\
&\leq \int \int_0^1 \left|\frac{\partial}{\partial x} I\left(\frac{i+t}{n} - x, \frac{j}{n} - y\right)\right| a(nx)b(ny) \\
&\leq \int \int_0^1 \left|\frac{\partial}{\partial x} I\left(\frac{t}{n} - x, -y\right)\right| a(nx+i)b(ny+j)
\end{aligned}$$

so that

$$\sum_{i,j} |I_{i+1,j} - I_{i,j}| \leq \sup_{x,y} \sum_{i,j} a(x+i)b(y+j) \int \left| \frac{\partial}{\partial x} I \right|,$$

and thus

$$||I_n||_{BV} \leq K \int \left| \frac{\partial}{\partial x} I \right| + \left| \frac{\partial}{\partial y} I \right| \leq \sqrt{2} K ||I||_{BV},$$

the last inequality resulting from $|u| + |v| \leq \sqrt{2}(u^2 + v^2)^{1/2}$. □

This theorem enables us to reformulate the fact that natural images do not belong to BV in a slightly different way. Suppose that the continuous image I is represented by the discrete function I_n of the previous theorem. Assume, moreover, that the distribution of the area of the discrete connected components of bilevels for I_n is $f_{n,\gamma}(k)$ for values of k from 1 to n^2. Then reasoning as in Theorem 4.5 leads to the following theorem.

THEOREM 5.2. *Let I be a function in $BV(C)$ and I_n a sampling of I, defined as in Lemma* 5.1*, the kernel G satisfying hypotheses* (5.3) *and* (5.4)*. Then there is a constant C such that*

$$||I||_{BV} \geq C \sum_{k=1}^{n^2} \left(\frac{1}{n^2} k \right)^{\frac{1}{2}} f_n(\gamma, k), \tag{5.5}$$

where $f_n(\gamma, k)$ is the number of connected components of I_n of area k for values of k from 1 *to n^2.*

Proof. Since I_n is a step function, all measurability conditions of Theorem 4.5 are satisfied, and we obtain formula (4.5), which yields

$$||I_n||_{BV} \geq C \sum_{k=1}^{n^2} \left(\frac{1}{n^2} k \right)^{\frac{1}{2}} f_n(\gamma, k).$$

By Lemma 5.1, we obtain formula (5.5). □

We now come to the consequences of our experiments.

COROLLARY 5.3. *If for all n there is a constant C_n such that*

$$f_n(\gamma, k) \geq \frac{C_n}{\left(\frac{1}{n^2} k\right)^{\alpha}} \tag{5.6}$$

with $\alpha < 2$, then

$$||I_n||_{BV} \geq C \sum_{k=1}^{n^2} \frac{1}{n^2} \left(\frac{k}{n^2} \right)^{\frac{1}{2} - \alpha}. \tag{5.7}$$

Proof. We have (computing the area of the unit square)

$$\sum_{k=1}^{n^2} f_n(\gamma, k) \frac{k}{n^2} = 1$$

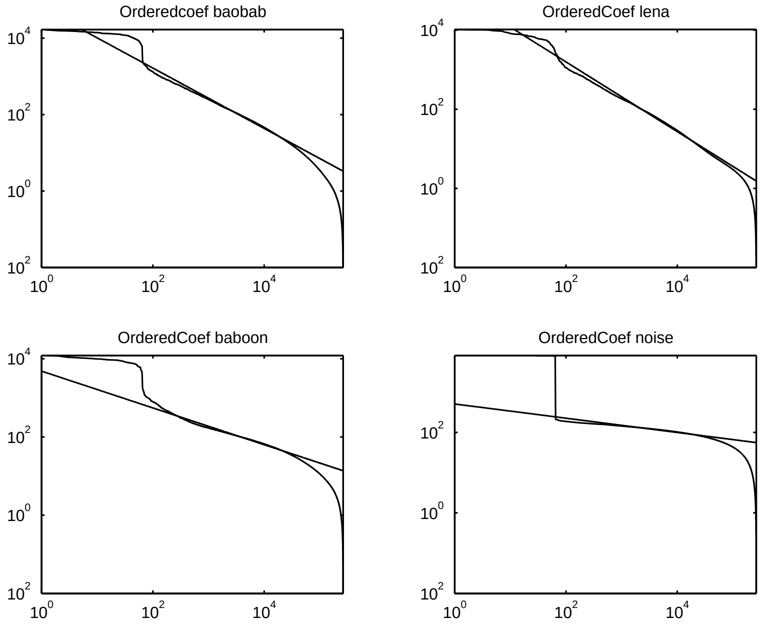


FIG. 3. *Ordered wavelet coefficients (modulus against the rank in* log-log *coordinates) and least squares fit for four images.*

so that

$$C_n \sum_{k=1}^{n^2} \frac{1}{n^2} \left(\frac{k}{n^2} \right)^{1-\alpha} \geq \frac{1}{n^2}.$$

Now if $\alpha < 2$, the preceding Riemann sum converges, and

$$C_n \geq \frac{(2-\alpha)}{n^2}.$$

Eventually, replacing expression (5.6) into formula (5.5) yields the result. □

Now the same conclusions as before hold, since the right side of formula (5.7) tends to infinity as soon as $\alpha > 1.5$. We have thus proved that if the distribution of the (discrete) sections of the (piecewise constant) function which is obtained by convolving I with a rescaled filter and sampling at n^2 points follows formula (5.6) with $1.5 < \alpha < 2$, then I is not of bounded variation.

Remark. The assumptions of Corollary 5.3 correspond exactly to our numerical experiments on natural images. Notice that we observed a variation of the constant C_n, and that the preceding proof shows that the blow-up result for the BV norm of images is independent of this variation.

6. Wavelets and the space BV. As mentioned in the introduction, recent results concerning the link between the global decay of wavelet coefficients and the total variation of images permit us to address the main question of this paper in a wavelet framework. Let (c_k) be the wavelet coefficients of the image I, ordered in a nonincreasing sequence. We say that the c_k's are in l^1 if $\sum |c_k| < +\infty$, and that they are in weak-l^1 if there exists a constant C such that $c_k \leq \frac{C}{k}$. Obviously, l^1 is included

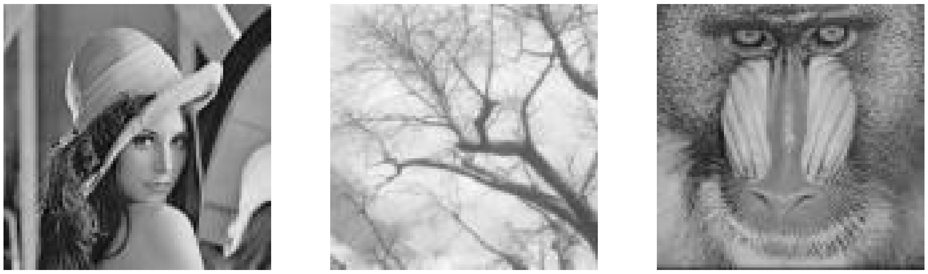


FIG. 4. *Lena, baobab, and baboon images, on which we study the decay of wavelet coefficients as shown in Figure* 3.

in weak-l^1. It is well known that if the c_k are in l^1, then I is in a Besov space included in BV. In the other direction, Cohen, DeVore, Petrushev, and Xu [CDPX99] recently proved that if I is in BV, then the c_k's are in weak-l^1 for the Haar wavelets. Cohen, Meyer, and Oru then generalized the result to any compactly supported wavelet basis; see [Oru98].

Thus it is possible to decide whether an image belongs to BV or not by looking at its wavelet coefficients decay, except if they decrease as $\frac{C}{k}$, which happens to be the case quite often. We present in Figure 3 the ordered coefficients for four images: the well-known Lena image, a part of a baobab image, a baboon image (Figure 4), and a Gaussian white noise. We used Daubechies's wavelets (using filters of length 8 provided by Wavelab), and found very similar results with Haar's wavelets. As we see, there is a fairly linear part in those graphs but only for intermediate scales. Assuming the small scales' behavior is perturbed by sampling and that the decay observed for most of the coefficients is characteristic of what happens at small scales, we may use the preceding results. We fitted a line to those values according to a least squares error so that an image should be in BV when the slope of this line is (strictly) greater than 1, and out of BV if the slope is (strictly) smaller than 1. In the case of the baobab and baboon images (Figure 4), we, respectively, found slopes of 0.76 and 0.45, and values of α (the exponent of the power law distribution of sections area), respectively, equal to 1.9 and 2.38 (for $k = 16$) so that both methods agree: those images are not in BV. For the well-known image of Lena, our approach gives an α of 1.9 (for $k = 16$), which suggests that Lena is clearly out of BV, whereas from the wavelet approach, a slope of .95 indicates the image is not BV. Now this result is close to the uncertainty zone. Of course, the decay of the coefficients for the noise image is very slow (0.16) so that this image is not in BV, whereas the distribution of the area of sections does not follow a power law (but an exponential distribution). Note that in all four cases, the inference of the distribution at small scales is less clear than in the morphological approach of the previous sections.

To understand the nature of the uncertainty when the slope is 1, it is worth noticing that the wavelet coefficients produced by the characteristic function of a simple shape already decrease as $\frac{1}{k}$ so that the simple presence of edges in the image implies this type of decay. In a sense, the wavelet coefficients look at the smoothness of the edges of the image, whereas, by investigating the sections' size distribution, we investigate the number and the cumulated length of those sections. Thus, as we mentioned already, it is not surprising that we get more precise estimates of the image oscillations at small scales and can therefore decide, for instance, that Lena is not BV while the wavelet coefficient analysis is ambiguous [Mal99b].

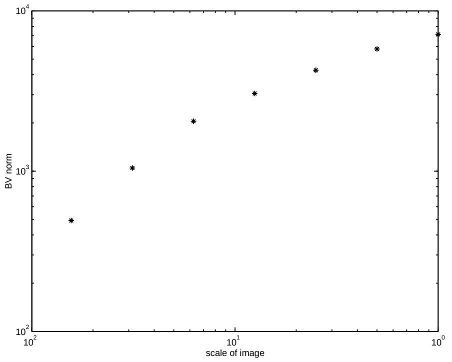


FIG. 5. *Simple plot of the BV norm as a function of the scale for the image of the airport. The BV norm is directly computed on the image which is successively downsampled. Abscissa equals a half means the image has been subsampled by a factor of* 2.

Clearly, our proposed procedure and the wavelet experiments imply that the scale behavior of the image needs some sophisticated statistics to be correctly extrapolated at fine scales. Thus it is necessary to point out that less sophisticated statistics can yield uninterpretable results on images for which the formerly mentioned methods yield clearcut answers. By one of the referee's suggestions, we performed the following experiment (Figure 5). We simulated a zoom backward at six dyadic scales of a natural image and computed the resulting BV norm. In several images where the section statistics are conclusive, we get no clear scaling behavior, as can be seen in Figure 5. Actually, even if the dots in the recently mentioned experiment had been aligned, we could not have made any strong statement. The number of obtainable samples is simply too small.

7. Conclusion. Combining experimental results about the distribution of sections in natural images and a result about the total variation of functions of $\mathbb{R}^2$, we have shown that natural images are not of bounded variation. This conclusion relies on the assumption that the observed images are obtained from continuous ones through fairly arbitrary convolution and sampling. This shows precisely that even if they are well-adapted to the large scale geometric structures of images, modeling them as functions with bounded variation does not account for the intricate nature of their small details.

REFERENCES

[AGM99] L. ALVAREZ, Y. GOUSSEAU, AND J.-M. MOREL, *The size of objects in natural and artificial images*, Advances in Imaging and Electron Physics, 111 (1999), pp. 167–242.

[Amb89] L. AMBROSIO, *Variational problems in SBV*, Acta Appl. Math., 17 (1989), pp. 1–40.

[ACMM99] L. AMBROSIO, V. CASELLES, S. MASNOU, AND J.-M. MOREL, *Connected components of sets of finite perimeter and applications to image processing*, J. Eur. Math. Soc. (JEMS), 3 (2001), pp. 39–92.

[ADK99] G. AUBERT, R. DERICHE, AND P. KORNPROBST, *Computing optical flow via variational techniques*, SIAM J. Appl. Math., 60 (1999), pp. 156–182.

[CW98] T. CHAN AND C. K. WONG, *Total variation blind deconvolution*, IEEE Trans. Image Process., 7 (1998), pp. 370–375.

[CDPX99] A. COHEN, R. DEVORE, P. PETRUSHEV, AND H. XU, *Non linear approximation and the space* $BV(\mathbb{R}^2)$, Amer. J. Math., 121 (1999), pp. 587–628.

[EG92] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC, Boca Raton, FL, 1992.

[Gou00] Y. Gousseau, *La distribution des formes dans les images naturelles*, Ph.D. thesis, CEREMADE, Université Paris IX, Paris, France, 2000.

[KMR99] J. Kalifa, S. Mallat, and B. Rougé, *Minimax deconvolution in mirror wavelet bases*, IEEE Trans. Image Process., submitted.

[Mal99a] F. Malgouyre, *Spectrum interpolation and image deblurring by means of the total variation*, IEEE Trans. Image Process., submitted.

[Mal99b] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed., Academic Press, New York, 1999.

[Mey93] Y. Meyer, *Wavelets: Algorithms and Applications*, SIAM, Philadelphia, 1993.

[Nik00] M. Nikolova, *Local strong homogeneity of a regularized estimator*, SIAM J. Appl. Math., 61 (2000), pp. 633–658.

[Oru98] F. Oru, *Rôle des oscillations dans quelques problèmes d'analyse non-linéaire*, Ph.D. thesis, ENS Cachan, Cachan, France, 1998.

[Rud87] L. Rudin, *Images, Numerical Analysis of Singularities and Shock Filters*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 1987.

[RO94] L. Rudin and S. Osher, *Total variation based image restoration with free local constraints*, in Proceedings of the IEEE International Conference on Image Processing, Vol. 1, Austin, Texas, 1994, pp. 31–35.

[ROF92] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.

[vHvdS98] J. van Hateren and A. van der Schaaf, *Independent component filters of natural images compared with simple cells in primary visual cortex*, Proc. Roy. Soc. London B, 265 (1998), pp. 359–366.

[Zie89] W. P. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, New York, 1989.

SIAM J. Math. Anal.
Vol. 33, No. 3, pp. 649–669

GLOBAL WELL-POSEDNESS FOR SCHRÖDINGER EQUATIONS WITH DERIVATIVE*

J. COLLIANDER[†], M. KEEL[‡], G. STAFFILANI[§], H. TAKAOKA[¶], AND T. TAO[‖]

Abstract. We prove that the one-dimensional Schrödinger equation with derivative in the nonlinear term is globally well-posed in H^s for $s > 2/3$, for small L^2 data. The result follows from an application of the "I-method." This method allows us to define a modification of the energy norm H^1 that is "almost conserved" and can be used to perform an iteration argument. We also remark that the same argument can be used to prove that any quintic nonlinear defocusing Schrödinger equation on the line is globally well-posed for large data in H^s, for $s > 2/3$.

Key words. Schrödinger equations, global well-posedness

AMS subject classifications. 35, 42

PII. S0036141001384387

1. Introduction.

We consider the derivative nonlinear Schrödinger initial value problem (IVP)

$$(1)\qquad \begin{cases} i\partial_t u + \partial_x^2 u = i\lambda\partial_x(|u|^2 u), \\ u(x,0) = u_0(x), \qquad\qquad x \in \mathbb{R}, \quad t \in \mathbb{R}, \end{cases}$$

where $\lambda \in \mathbb{R}$. The equation in (1) is a model for the propagation of circularly polarized Alfvén waves in magnetized plasma with a constant magnetic field [18, 19, 23].

It is natural to impose the smallness condition

$$(2)\qquad \|u_0\|_{L^2} < \sqrt{\frac{2\pi}{\lambda}}$$

on the initial data, as this will force the energy to be positive via the sharp Gagliardo–Nirenberg inequality. Note that the L^2 norm is conserved by the evolution.

Well-posedness for the Cauchy problem (1) has been studied by many authors [10, 11, 12, 20, 21, 22, 26, 27]. The best local well-posedness result is due to Takaoka [22], who used a gauge transformation and the Fourier restriction method is used to obtain local well-posedness in $H^s,\ s \geq 1/2$. In [24], Takaoka showed that this result is sharp in the sense that the data map fails to be C^3 or uniformly C^0 for $s < 1/2$ (cf. Bourgain [4] and Biagioni–Linares [1]).

*Received by the editors January 31, 2001; accepted for publication (in revised form) June 11, 2001; published electronically October 4, 2001.

http://www.siam.org/journals/sima/33-3/38438.html

[†] Department of Mathematics, University of Toronto, Toronto, Canada M5S3G3 (colliand@math.toronto.edu). This author's work was supported in part by an NSF Postdoctoral Research Fellowship.

[‡] School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (keel@math.umn.edu). This author's work was supported in part by NSF grant DMS 9801558.

[§] Department of Mathematics, Brown University, Providence, RI 02912 (gigliola@math.brown.edu). This author's work was supported in part by NSF grant DMS 9800879 and a grant from Hewlett-Packard.

[¶] Department of Mathematics, Hokkaido University, Sapporo, 060-0810 Japan (takaoka@math.sci.hokudai.ac.jp). This author's work was supported in part by JSPS grant 13740087.

[‖] Department of Mathematics, University of California Los Angeles, Los Angeles, CA 90095 (tao@math.ucla.edu). This author is a Clay Prize Fellow and was supported in part by grants from the Packard and Sloan Foundations.

In [20], global well-posedness is obtained for (1) in H^1, assuming the smallness condition (2). The argument there is based on two gauge transformations performed in order to remove the derivative in the nonlinear term. This was improved by Takaoka [24], who proved global well-posedness in H^s for $s > \frac{32}{33}$, assuming (2). The method of proof is based on the idea of Bourgain [3, 5] of estimating separately the evolution of low frequencies and of high frequencies of the initial data.

In this paper we improve the global well-posedness result further.

Theorem 1.1. *The Cauchy problem* (1) *is globally well-posed in* H^s *for* $s > 2/3$, *assuming the smallness condition* (2).

The proof of Theorem 1.1 is based on the "I-method" used by the authors in other nonlinear Cauchy problems in [15, 7, 8, 9] (see also [14]). The basic idea is as follows. After a rescaling, we define a new energy $E_N(u)(t)$ for the solution u that depends on a parameter $N \gg 1$. We prove a local well-posed result in the norm associated to E_N on intervals of length ~ 1, and finally we perform an iteration on the time intervals. The reason why this iteration can be globally extended is that the increment of the energy $E_N(u)(t)$ over each time interval is very small. In other words, the argument is successful because the energy $E_N(u)(t)$ is *almost conserved*.

After the proof of Theorem 1.1 is completed, we will briefly remark that, using the same techniques, one can also show that the one-dimensional defocusing quintic nonlinear Schrödinger is globally well-posed for initial data in $H^s, s > 2/3$. The details of the proof of this fact will appear in a different paper.

The restriction $s > 2/3$ is probably not sharp and might be improvable either by more sophisticated multilinear estimates and better estimates on the symbols M_4, M_6, M_8 which appear in our argument, or by using the "correction term" strategy of [8]. In fact, one may reasonably conjecture that one could extend the global well-posedness result to match the local result at $s > 1/2$. We will not pursue these matters here.

2. Notation.

To prove Theorem 1.1, we may assume $2/3 < s < 1$, since for $s \geq 1$ the result is contained in [20, 24]. Henceforth $2/3 < s < 1$ shall be fixed. Also, by rescaling u, we may assume $\lambda = 1$.

We use C to denote various constants depending on s; if C depends on other quantities as well, this will be indicated by explicit subscripting, e.g., $C_{\|u_0\|_2}$ will depend on both s and $\|u_0\|_2$. We use $A \lesssim B$ to denote an estimate of the form $A \leq CB$. We use $a+$ and $a-$ to denote expressions of the form $a + \varepsilon$ and $a - \varepsilon$, where $0 < \varepsilon \ll 1$ depends only on s.

We use $\|f\|_p$ to denote the $L^p(\mathbb{R})$ norm and $L^q_t L^r_x$ to denote the mixed norm

$$\|f\|_{L^q_t L^r_x} := \left(\int \|f(t)\|_r^q \, dt \right)^{1/q}$$

with the usual modifications when $q = \infty$.

We define the spatial Fourier transform of $f(x)$ by

$$\hat{f}(\xi) := \int_{\mathbb{R}} e^{-ix\xi} f(x) \, dx$$

and the spacetime Fourier transform $u(t, x)$ by

$$\tilde{u}(\tau, \xi) := \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-i(x\xi + t\tau)} u(t, x) \, dt dx.$$

Note that the derivative ∂_x is conjugated to multiplication by $i\xi$ by the Fourier transform.

We shall also define D_x to be the Fourier multiplier with symbol $\langle\xi\rangle := 1 + |\xi|$. We can then define the Sobolev norms H^s by

$$\|f\|_{H^s} := \|D_x^s f\|_2 = \|\langle\xi\rangle^s \hat{f}\|_{L^2_\xi}.$$

We also define the spaces $X^{s,b}(\mathbb{R} \times \mathbb{R})$ (first introduced in [2]) on $\mathbb{R} \times \mathbb{R}$ by

$$\|u\|_{X^{s,b}(\mathbb{R}\times\mathbb{R})} := \|\langle\xi\rangle^s \langle\tau - |\xi|^2\rangle^b \hat{u}(\xi,\tau)\|_{L^2_\tau L^2_\xi}.$$

We often use $\|u\|_{s,b}$ to abbreviate $\|u\|_{X^{s,b}(\mathbb{R}\times\mathbb{R})}$. For any time interval I, we define the restricted spaces $X^{s,b}(I \times \mathbb{R})$ by

$$\|u\|_{X^{s,b}(I\times\mathbb{R})} := \inf\{\|U\|_{s,b} : U|_{I\times\mathbb{R}} = u\}.$$

We shall take advantage of the Strichartz estimates

$$\|u\|_{L^6_t L^6_x} \lesssim \|u\|_{0,1/2+} \tag{3}$$

and

$$\|u\|_{L^\infty_t L^2_x} \lesssim \|u\|_{0,1/2+} \tag{4}$$

(see, e.g., [2]). From (4) and Sobolev embedding we observe

$$\|u\|_{L^\infty_t L^\infty_x} \lesssim \|u\|_{1/2+,1/2+}. \tag{5}$$

In our arguments we shall be using the trivial embedding

$$\|u\|_{s_1,b_1} \lesssim \|u\|_{s_2,b_2} \text{ whenever } s_1 \leq s_2, b_1 \leq b_2$$

so frequently that we will not mention this embedding explicitly.

We now give some useful notation for multilinear expressions. If $n \geq 2$ is an even integer, we define a *(spatial) multiplier of order* n to be any function $M_n(\xi_1, \ldots, \xi_n)$ on the hyperplane

$$\Gamma_n := \{(\xi_1, \ldots, \xi_n) \in \mathbb{R}^n : \xi_1 + \cdots + \xi_n = 0\},$$

which we endow with the standard measure $\delta(\xi_1 + \cdots + \xi_n)$, where δ is the Dirac delta.

If M_n is a multiplier of order n and $f_1, \ldots, f_n$ are functions on $\mathbb{R}$, we define the quantity $\Lambda_n(M_n; f_1, \ldots, f_n)$ by

$$\Lambda_n(M_n; f_1, \ldots, f_n) := \int_{\Gamma_n} M_n(\xi_1, \ldots, \xi_n) \prod_{j=1}^{n} \hat{f}_j(\xi_j).$$

We adopt the notation

$$\Lambda_n(M_n; f) := \Lambda_n(M_n; f, \bar{f}, f, \bar{f}, \ldots, f, \bar{f}).$$

Observe that $\Lambda_n(M_n; f)$ is invariant under permutations of the even ξ_j indices or of the odd ξ_j indices.

If M_n is a multiplier of order n, $1 \le j \le n$ is an index, and $k \ge 1$ is an even integer, we define the *elongation* $\mathbf{X}_j^k(M_n)$ of M_n to be the multiplier of order $n+k$ given by

$$\mathbf{X}_j^k(M_n)(\xi_1, \dots, \xi_{n+k}) := M_n(\xi_1, \dots, \xi_{j-1}, \xi_j + \cdots + \xi_{j+k}, \xi_{j+k+1}, \dots, \xi_{n+k}).$$

In other words, $\mathbf{X}_j^k$ is the multiplier obtained by replacing ξ_j by $\xi_j + \cdots + \xi_{j+k}$ and advancing all the indices after ξ_j accordingly.

We shall often write ξ_{ij} for $\xi_i + \xi_j$, ξ_{ijk} for $\xi_i + \xi_j + \xi_k$, etc. We also write ξ_{i-j} for $\xi_i - \xi_j$, ξ_{ij-klm} for $\xi_{ij} - \xi_{klm}$, etc.

3. The Gauge transformation and the conservation laws. In this section we apply the gauge transform used in [20] in order to improve the derivative nonlinearity.

DEFINITION 3.1. *We define the nonlinear map* $\mathcal{G} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ *by*

$$\mathcal{G}f(x) := e^{-i\int_{-\infty}^{x} |f(y)|^2 dy} f(x).$$

The inverse transform $\mathcal{G}^{-1}f$ *is then given by*

$$\mathcal{G}^{-1}f(x) := e^{i\int_{-\infty}^{x} |f(y)|^2 dy} f(x).$$

This transform is well behaved on H^s.

LEMMA 3.2. *The map* $\mathcal{G}$ *is a bicontinuous map from* H^s *to* H^s.

A similar statement holds for $0 \le s \le 1/2$, but we shall not need it here.

Proof. We shall just prove the continuity of $\mathcal{G}$, as the continuity of $\mathcal{G}^{-1}$ is proven similarly.

Define *Lip* to be the space of functions with norm

$$\|f\|_{Lip} := \|f\|_\infty + \|f'\|_{L^\infty}.$$

Since $s > 1/2$, we see from Sobolev embedding that the nonlinear map $f \mapsto e^{-i\int_{-\infty}^{x} |f(y)|^2 dy}$ continuously maps H^s to *Lip*. It therefore suffices to show the product estimate

$$\|fg\|_{H^s} \lesssim \|f\|_{H^s} \|g\|_{Lip}.$$

But this estimate follows immediately from the Leibniz rule and Hölder when $s = 0$ or $s = 1$, and the intermediate cases then follow by interpolation. □

Set $w_0 := \mathcal{G}u_0$ and $w(t) := \mathcal{G}u(t)$ for all times t. A straightforward calculation shows that the IVP (1) can be transformed to

$$\begin{cases} i\partial_t w + \partial_x^2 w = -iw^2 \partial_x \bar{w} - \frac{1}{2}|w|^4 w, \\ w(x,0) = w_0(x), \qquad x \in \mathbb{R}, \quad t \in \mathbb{R}. \end{cases} \tag{6}$$

Also, the smallness condition (2) becomes

$$\|w_0\|_{L^2} < \sqrt{2\pi}. \tag{7}$$

By Lemma 3.2 we thus see that global well-posedness of (1) in H^s is equivalent to that of (6). From [20, 22, 24], we know that both Cauchy problems are locally well-posed in H^s and globally well-posed in H^1, assuming (7). By standard limiting arguments, we thus see that Theorem 1.1 will follow if we can show the following.

Proposition 3.3. *Let w be a global H^1 solution to* (6) *obeying* (7). *Then for any $T > 0$ we have*

$$\sup_{0\leq t\leq T} \|w(t)\|_{H^s} \lesssim C_{\|w_0\|_{H^s}, \|w_0\|_2, T},$$

where the right-hand side does not depend on the H^1 norm of w.

Just by looking at the equation in (6) it is not easy to understand why this should be better than the equation in (1). In fact, we still see a derivative, and, moreover, a quintic nonlinearity has been introduced. But it was made clear in [20, 13, 22] how a derivative of the complex conjugate of the solution w can be handled while a derivative of w cannot. Also, the quintic term is not going to introduce any extra trouble.

Let $n \geq 2$ be an even integer, and let M_n be a multiplier of order n. From (6) we have

$$\partial_t w = i w_{xx} - w\bar{w}_x w + \frac{i}{2} w\bar{w}w\bar{w}w$$

and

$$\partial_t \bar{w} = -i w_{xx} - \bar{w} w_x \bar{w} - \frac{i}{2} \bar{w}w\bar{w}w\bar{w}.$$

Taking the Fourier transform of these identities, we obtain the useful differentiation law

$$\begin{aligned}\partial_t \Lambda_n(M_n; w(t)) = {} & i\Lambda_n\left(M_n \sum_{j=1}^{n} (-1)^j \xi_j^2; w(t)\right)\\ & - i\Lambda_{n+2}\left(\sum_{j=1}^{n} \mathbf{X}_j^2(M_n)\xi_{j+1}; w(t)\right)\\ & + \frac{i}{2}\Lambda_{n+4}\left(\sum_{j=1}^{n} (-1)^{j-1}\mathbf{X}_j^4(M_n); w(t)\right)\end{aligned} \tag{8}$$

for any even integer $n \geq 2$ and any multiplier M_n of order n.

We now turn to the conservation laws that the solution w of (6) enjoys. What follows in this section was originally described by Ozawa in [20]; however, we have redone the computations in our own notation, as this will prove useful later.

Definition 3.4. *If $f \in H^1(\mathbb{R})$, we define the energy $E(f)$ by*

$$E(f) := \int \partial_x f \partial_x \overline{f}\; dx - \frac{1}{2}\mathrm{Im}\int f\overline{f} f \partial_x \overline{f}\; dx.$$

By Plancherel, we may write $E(f)$ using the Λ notation as

$$E(f) = -\Lambda_2(\xi_1\xi_2; f) - \frac{1}{2}\mathrm{Im}\Lambda_4(i\xi_4; f).$$

Expanding the second term using $\mathrm{Im}(z) = (z - \bar{z})/2i$ and using symmetry, we may rewrite this as

$$E(f) = -\Lambda_2(\xi_1\xi_2; f) + \frac{1}{8}\Lambda_4(\xi_{13-24}; f). \tag{9}$$

Lemma 3.5 (see [20]). *If w is an H^1 solution to (6) for times $t \in [0,T]$, then we have*

$$\|w(t)\|_2 = \|w_0\|_2$$

and

$$E(w(t)) = E(w_0)$$

for all $t \in [0,T]$.

Proof. These conservation laws are proven in [20]; however, we give a proof based on the identity (8), as the proof here will be needed later on.

We of course have

$$\|w(t)\|_2^2 = \Lambda_2(1; w(t)).$$

In the rest of this proof we shall drop the $w(t)$ from the Λ notation. Differentiating the previous equation and applying (8), we obtain

$$\partial_t \|w(t)\|_2^2 = -i\Lambda_2(\xi_1^2 - \xi_2^2) - i\Lambda_4(\xi_2 + \xi_3) + \frac{i}{2}\Lambda_6(1 - 1 + 1 - 1 + 1 - 1).$$

The first term vanishes since $\xi_{12} = 0$. The second term can be symmetrized to $-\frac{i}{2}\Lambda_4(\xi_{1234})$, which vanishes. The third term clearly vanishes. This proves the L^2 conservation.

Now we prove energy conservation. From (9) we have

$$\partial_t E(t) = -\partial_t \Lambda_2(\xi_1\xi_2) + \frac{1}{8}\partial_t \Lambda_4(\xi_{13-24}), \tag{10}$$

and from (8) we have

$$\partial_t \Lambda_2(\xi_1\xi_2) = -i\Lambda_2(\xi_1\xi_2(\xi_1^2 - \xi_2^2)) - i\Lambda_4(\xi_{123}\xi_4\xi_2 + \xi_1\xi_{234}\xi_3) + \frac{i}{2}\Lambda_6(\xi_{12345}\xi_6 - \xi_1\xi_{23456}).$$

The Λ_2 term vanishes since $\xi_{12} = 0$. To simplify the Λ_4 term we write $\xi_{123} = -\xi_4$ and $\xi_{234} = -\xi_1$ and then symmetrize. To simplify the Λ_6 term we write $\xi_{12345} = -\xi_6$ and $\xi_{23456} = -\xi_1$ and then symmetrize to obtain

$$\partial_t \Lambda_2(\xi_1\xi_2) = \frac{i}{2}\Lambda_4(\xi_1^2\xi_3 + \xi_2^2\xi_4 + \xi_3^2\xi_1 + \xi_4^2\xi_2) + \frac{i}{6}\Lambda_6(\xi_1^2 - \xi_2^2 + \xi_3^2 - \xi_4^2 + \xi_5^2 - \xi_6^2).$$

We may simplify the Λ_4 term further, using the identity

$$\begin{aligned}
\xi_1^2\xi_3 + \xi_2^2\xi_4 + \xi_3^2\xi_1 + \xi_4^2\xi_2 &= \xi_1\xi_3\xi_{13} + \xi_2\xi_4\xi_{24} \\
&= \xi_{13}(\xi_1\xi_3 - \xi_2\xi_4) \\
&= \xi_{13}(-\xi_1\xi_{124} - \xi_2\xi_4) \\
&= -\xi_{13}(\xi_1 + \xi_2)(\xi_1 + \xi_4) \\
&= -\xi_{12}\xi_{13}\xi_{14}
\end{aligned}$$

to obtain

$$\partial_t \Lambda_2(\xi_1\xi_2) = -\frac{i}{2}\Lambda_4(\xi_{12}\xi_{13}\xi_{14}) + \frac{i}{6}\Lambda_6(\xi_1^2 - \xi_2^2 + \xi_3^2 - \xi_4^2 + \xi_5^2 - \xi_6^2). \tag{11}$$

We now consider the second component of the energy. From (8) we have

$$\begin{aligned}\partial_t \Lambda_4(\xi_{13-24}) &= i\Lambda_4(\xi_{13-24}(\xi_1^2 - \xi_2^2 + \xi_3^2 - \xi_4^2)) \\ &\quad - i\Lambda_6(\xi_{1235-46}\xi_2 + \xi_{15-2346}\xi_3 + \xi_{1345-26}\xi_4 + \xi_{13-2456}\xi_5) \\ &\quad + \frac{i}{2}\Lambda_8(\xi_{123457-68} - \xi_{17-234568} + \xi_{134567-28} - \xi_{13-245678}).\end{aligned}$$

The Λ_8 term symmetrizes to $i\Lambda_8(\xi_{12345678})$, which vanishes. The Λ_6 term can be rewritten as

$$2i\Lambda_6(\xi_{46}\xi_2 - \xi_{15}\xi_3 + \xi_{26}\xi_4 - \xi_{13}\xi_5),$$

which we rewrite as

$$2i\Lambda_6(\xi_{246}\xi_2 - \xi_{135}\xi_3 + \xi_{246}\xi_4 - \xi_{135}\xi_5) - 2i\Lambda_6(\xi_2^2 - \xi_3^2 + \xi_4^2 - \xi_5^2).$$

The first term symmetrizes to $\frac{4i}{3}\Lambda_6(\xi_{246}^2 - \xi_{135}^2)$, which vanishes. The second term symmetrizes to

$$\frac{4i}{3}\Lambda_6(\xi_1^2 - \xi_2^2 + \xi_3^2 - \xi_4^2 + \xi_5^2 - \xi_6^2).$$

Finally, consider the Λ_4 term. We may factorize

$$\xi_{13-24}(\xi_1^2 - \xi_2^2 + \xi_3^2 - \xi_4^2) = \xi_{13-24}(\xi_{1-2}\xi_{12} + \xi_{3-4}\xi_{34}).$$

Since $\xi_{12} = \xi_{-34}$ and $\xi_{13} = -\xi_{24}$, we may simplify this as

$$2\xi_{13}\xi_{12}(\xi_{1-2} - \xi_{3-4}) = 4\xi_{12}\xi_{13}\xi_{14}.$$

Combining all these identities, we thus have

$$\frac{1}{8}\partial_t \Lambda_4(\xi_{13-24}) = -\frac{1}{2}i\Lambda_4(\xi_{12}\xi_{13}\xi_{14}) - \frac{i}{6}\Lambda_6(\xi_1^2 - \xi_2^2 + \xi_3^2 - \xi_4^2 + \xi_5^2 - \xi_6^2).$$

Combining this with (11) and (10), we obtain

$$\partial_t E(w(t)) = 0,$$

and the claim follows. □

Heuristically, the energy $E(w(t))$ has the same strength as $\|w(t)\|_{H^1}^2$. We can make this precise as follows.

Lemma 3.6. *Let f be an H^1 function on $\mathbb{R}$ such that $\|f\|_2 < \sqrt{2\pi}$. Then we have*

$$\|\partial_x f\|_2 \le C_{\|f\|_2} E(f)^{1/2}, \tag{12}$$

where $C_{\|f\|_2}$ depends only on $\|f\|_2$.

Proof. Define the function

$$g(x) := \exp\left(i\frac{3}{4}\int_{-\infty}^{x} |f(y)|^2\, dy\right) f(x).$$

A routine computation shows that

$$\|g\|_2 = \|f\|_2 < \sqrt{2\pi}$$

and

$$E(f) = \|\partial_x g\|_2^2 - \frac{1}{16}\|g\|_6^6.$$

From the sharp Gagliardo–Nirenberg inequality [28]

$$\|g\|_6^6 \leq \frac{4}{\pi^2}\|g\|_2^4\|\partial_x g\|_2^2, \tag{13}$$

we therefore have

$$\|\partial_x g\|_2 \lesssim C_{\|f\|_2} E(f)^{1/2}.$$

From the definition of g we have

$$f(x) = \exp\left(-i\frac{3}{4}\int_{-\infty}^{x} |g(y)|^2 \; dy\right) g(x),$$

and so we have

$$\|\partial_x f\|_2 \lesssim \|\partial_x g\|_2 + \|g^3\|_2.$$

By another application of (13) we thus obtain (12). □

4. The almost conserved energy norm. It remains to prove Proposition 3.3. Fix w, T. We also let $N \gg 1$ be a large parameter depending on T, $\|w_0\|_2$, and $\|w_0\|_{H^s}$ which we shall choose later.

Because we do not want to use the H^1 norm of w, we cannot directly use the energy $E(w(t))$ defined above. So we are looking for a substitute notion of "energy" that can be defined for a less regular solution and that has a very slow increment in time. In the frequency space let us consider an even C^∞ monotone multiplier $m(\xi)$ taking values in $[0, 1]$ such that

$$m(\xi) := \begin{cases} 1 \quad \text{if } |\xi| < N, \\ \left(\frac{|\xi|}{N}\right)^{s-1} \quad \text{if } |\xi| > 2N. \end{cases} \tag{14}$$

We define the multiplier operator $I : H^s \longrightarrow H^1$ such that $\widehat{Iw}(\xi) := m(\xi)\widehat{w}(\xi)$. This operator is smoothing of order $1 - s$; indeed, we have

$$\|u\|_{s_0,b_0} \lesssim \|Iu\|_{s_0+1-s,b_0} \lesssim N^{1-s}\|u\|_{s_0,b_0} \tag{15}$$

for any $s_0, b_0 \in \mathbb{R}$.

Our substitute energy will be defined by

$$E_N(w) := E(Iw).$$

Note that this energy makes sense even if w is in only H^s.

In general the energy $E_N(w(t))$ is not conserved in time, but we will show that the increment is very small in terms of N. This will be accomplished in three stages. First, in Proposition 4.1 below, we write the increment of $E_N(w(t))$ as a multilinear expression in w. Then, in Lemma 6.1, we estimate these multilinear expressions in terms of the norm $\|Iw\|_{1,1/2+}$, gaining a power of N^{-1+} in the process. Finally, in

Theorem 5.1 (and Lemma 3.6), we control the norm $\|Iw\|_{1,1/2+}$ back in terms of $E_N(w(t))$.

Proposition 4.1. *Let w be an H^1 global solution to* (6). *Then for any $T \in \mathbb{R}$ and $\delta > 0$ we have*

$$E_N(w(T+\delta)) - E_N(w(T)) = \int_T^{T+\delta} [\Lambda_4(M_4; w(t)) + \Lambda_6(M_6; w(t)) + \Lambda_8(M_8; w(t))]\, dt,$$

where the multipliers M_4, M_6, M_8 are given by

$$\begin{aligned} M_4 &:= C_1 m_1 m_2 m_3 m_4 \xi_{12}\xi_{13}\xi_{14} + C_2(m_1^2\xi_1^2\xi_3 + m_2^2\xi_2^2\xi_4 + m_3^2\xi_3^2\xi_1 + m_4^2\xi_4^2\xi_2),\\ M_6 &:= C_3 \sum_{j=1}^{6} (-1)^{j-1} m_j^2 \xi_j^2 + C_4 \sum_{\{a,c,e\}=\{1,3,5\},\{b,d,f\}=\{2,4,6\}} m_a m_b m_c m_{def} \xi_{ac} \xi_e\\ &\quad - \ m_{abc} m_d m_e m_f \xi_{df} \xi_b,\\ M_8 &:= C_5 \sum_{\{a,c,e,g\}=\{1,3,5,7\};\{b,d,f,h\}=\{2,4,6,8\}} m_a m_b m_c m_{defgh} \xi_{ac-bdefgh}\\ &\quad - \ m_{abcde} m_f m_g m_h \xi_{abcdeg-fh}, \end{aligned}$$

where $C_1, \dots, C_5$ are absolute constants and we adopt the abbreviations m_i for $m(\xi_i)$, m_{ij} for $m(\xi_{ij})$, etc. Furthermore, if $|\xi_j| \ll N$ for all j, then the multipliers M_4, M_6, M_8 all vanish.

Proof. From (9) we have

$$E_N(w(t)) = -\Lambda_2(m_1\xi_1 m_2\xi_2; w(t)) + \frac{1}{8}\Lambda_4(\xi_{13-24} m_1 m_2 m_3 m_4; w(t)).$$

Henceforth we omit the $w(t)$ from the Λ notation. By (8) we have

$$\begin{aligned} \partial_t \Lambda_2(m_1\xi_1 m_2\xi_2) &= -i\Lambda_2(m_1\xi_1 m_2\xi_2(\xi_1^2 - \xi_2^2))\\ &\quad - i\Lambda_4(m_{123}\xi_{123} m_4\xi_4\xi_2 + m_1\xi_1 m_{234}\xi_{234}\xi_3)\\ &\quad + \frac{i}{2}\Lambda_6(m_{12345}\xi_{12345} m_6\xi_6 - m_1\xi_1 m_{23456}\xi_{23456}). \end{aligned}$$

The Λ_2 term vanishes since $\xi_{12} = 0$. To simplify the Λ_4 term, we use $\xi_{123} = -\xi_4$ and $\xi_{234} = -\xi_1$ and then symmetrize to obtain the second term of M_4. To simplify the Λ_6 term, we use $\xi_{12345} = -\xi_6$ and $\xi_{23456} = -\xi_1$ and then symmetrize to get the first term of M_6.

In a similar vein, we have

$$\begin{aligned} \partial_t \Lambda_4(\xi_{13-24} m_1 m_2 m_3 m_4) &= -i\Lambda_4(\xi_{13-24} m_1 m_2 m_3 m_4 (\xi_1^2 - \xi_2^2 + \xi_3^2 - \xi_4^2))\\ &\quad - i\Lambda_6(\xi_{1235-46} m_{123} m_4 m_5 m_6 \xi_2 + \xi_{15-2346} m_1 m_{234} m_5 m_6 \xi_3\\ &\qquad + \xi_{1345-26} m_1 m_2 m_{345} m_6 \xi_4 + \xi_{13-2456} m_1 m_2 m_3 m_{456} \xi_5)\\ &\quad + \frac{i}{2}\Lambda_8(\xi_{123457-68} m_{12345} m_6 m_7 m_8 - \xi_{17-234568} m_1 m_{23456} m_7 m_8\\ &\qquad + \xi_{134567-28} m_1 m_2 m_{34567} m_8 - \xi_{13-245678} m_1 m_2 m_3 m_{45678}). \end{aligned}$$

The Λ_4 term is of the form of the first term of M_4 by the argument used to prove (11). To simplify the Λ_6 term, we use $\xi_{1235-46} = -2\xi_{46}$ and similarly for the other four terms and then symmetrize to obtain the second term of M_6. Finally, if we symmetrize the Λ_8 term, we obtain M_8. The first part of the proposition then follows from the fundamental theorem of calculus applied to the function $t \longrightarrow E_N(w(t))$.

If all the frequencies are $\ll N$, then all the m_i, m_{ij}, etc. terms are equal to 1. In this case our calculations are identical to those in Lemma 3.5, and so our symbols M_4, M_6, M_8 will vanish by the computations given in that lemma. □

5. Local estimates. In Lemma 6.1 we shall estimate the expression in Proposition 4.1. It turns out that one cannot estimate this expression effectively just by using spatial norms such as $\|Iw\|_{H^1}$ (as is done for some simple equations in [5]), but one must use spacetime norms such as $\|Iw\|_{1,1/2+}$. The purpose of this section is to obtain the required control on these spacetime norms.

Theorem 5.1. *Let w be an H^1 global solution to* (6), *and let $T \in \mathbb{R}$ be such that*

$$\|Iw(T)\|_{H^1} \leq C_0$$

for some $C_0 > 0$. Then we have

$$\|Iw\|_{X^{1,1/2+}([T,T+\delta]\times\mathbb{R})} \lesssim 1$$

for some $\delta > 0$ depending on C_0.

We now prove Theorem 5.1. We shall be able to exploit the estimates in [22]. By standard iteration arguments (see, e.g., [2, 16, 17, 22, 24, 25]) it suffices to prove the following lemma.

Lemma 5.2. *We have*

$$\|I(w_1\partial_x\overline{w_2}w_3)\|_{X_{1,b-1}(\mathbb{R}\times\mathbb{R})} \lesssim \prod_{i=1}^{3}\|Iw_i\|_{X_{1,1/2+}(\mathbb{R}\times\mathbb{R})}, \tag{16}$$

$$\|I(w_1\overline{w_2}w_3\overline{w_4}w_5)\|_{X_{1,b-1}(\mathbb{R}\times\mathbb{R})} \lesssim \prod_{i=1}^{5}\|Iw_i\|_{X_{1,1/2+}(\mathbb{R}\times\mathbb{R})} \tag{17}$$

for all Schwarz functions w_i and some $b > 1/2$ (in fact, we may take any $1/2 < b < 5/8$).

Proof. By Plancherel and duality it suffices to show

$$\left|\int_* \frac{m(\xi_4)\langle\xi_4\rangle\langle\tau_4+\xi_4^2\rangle^{b-1}\xi_2}{\prod_{j=1}^{3} m(\xi_j)\langle\xi_j\rangle\langle\tau_j-(-1)^{j-1}\xi_j^2\rangle^{1/2+}}\prod_{j=1}^{4}F_j(\tau_j,\xi_j)\right| \lesssim \prod_{j=1}^{4}\|F_j\|_{L^2_{\tau_j}L^2_{\xi_j}}$$

and

$$\left|\int_{**} \frac{m(\xi_6)\langle\xi_6\rangle\langle\tau_6+\xi_6^2\rangle^{b-1}}{\prod_{j=1}^{5} m(\xi_j)\langle\xi_j\rangle\langle\tau_j-(-1)^{j-1}\xi_j^2\rangle^{1/2+}}\prod_{j=1}^{6}F_j(\tau_j,\xi_j)\right| \lesssim \prod_{j=1}^{6}\|F_j\|_{L^2_{\tau_j}L^2_{\xi_j}}$$

for all functions $F_1, \ldots, F_6$, where $\int_*$, $\int_{**}$ denote integration over the measure $\delta(\tau_1 + \cdots + \tau_4)\delta(\xi_1 + \cdots + \xi_4)$ and $\delta(\tau_1 + \cdots + \tau_6)\delta(\xi_1 + \cdots + \xi_6)$, respectively.

We may assume that the F_j are all real and nonnegative. We now observe the pointwise estimate

$$\frac{m(\xi_n)\langle\xi_n\rangle^{1-s}}{\prod_{j=1}^{n-1} m(\xi_j)\langle\xi_j\rangle^{1-s}} \lesssim 1$$

for $n = 4, 6$ and all $\xi_1, \dots, \xi_n$ such that $\xi_1 + \cdots + \xi_n = 0$. To see this, we use symmetry to assume that $|\xi_1| \geq \cdots \geq |\xi_{n-1}|$ so that $|\xi_n| \lesssim |\xi_1|$. Since $m(\xi)\langle\xi\rangle^{1-s}$ is essentially increasing in $|\xi|$, we thus see that

$$\frac{m(\xi_n)\langle\xi_n\rangle^{1-s}}{\prod_{j=1}^{n-1} m(\xi_j)\langle\xi_j\rangle^{1-s}} \lesssim \frac{1}{\prod_{j=2}^{n-1} m(\xi_j)\langle\xi_j\rangle^{1-s}}.$$

Since $m(\xi)\langle\xi\rangle^{1-s} \gtrsim 1$ for all ξ, the claim follows.

Because of this estimate, we need only to show the estimates

$$\left|\int_* \frac{\langle\xi_4\rangle^s\langle\tau_4 + \xi_4^2\rangle^{b-1}\xi_2}{\prod_{j=1}^3 \langle\xi_j\rangle^s\langle\tau_j - (-1)^{j-1}\xi_j^2\rangle^{1/2+}} \prod_{j=1}^4 F_j(\tau_j, \xi_j)\right| \lesssim \prod_{j=1}^4 \|F_j\|_{L^2_{\tau_j} L^2_{\xi_j}} \tag{18}$$

and

$$\left|\int_{**} \frac{\langle\xi_6\rangle^s\langle\tau_6 + \xi_6^2\rangle^{b-1}}{\prod_{j=1}^5 \langle\xi_j\rangle^s\langle\tau_j - (-1)^{j-1}\xi_j^2\rangle^{1/2+}} \prod_{j=1}^6 F_j(\tau_j, \xi_j)\right| \lesssim \prod_{j=1}^6 \|F_j\|_{L^2_{\tau_j} L^2_{\xi_j}}. \tag{19}$$

The estimate (18) is equivalent to the first estimate of Lemma 3.1 in [24] after undoing the duality and Plancherel, so it suffices to prove (19). By undoing the duality we can write this as

$$\|w_1\overline{w_2}w_3\overline{w_4}w_5\|_{s,b-1} \lesssim \prod_{j=1}^5 \|w_j\|_{s,1/2+}.$$

We may assume that the Fourier transforms $\tilde{w}_j$ are all real and nonnegative. By using $|\xi_1 + \xi_2 + \xi_3 + \xi_4 + \xi_5|^s \lesssim \sum_{i=1}^5 |\xi_i|^s$, it suffices to prove estimates of the form

$$\|(D_x^s w_1)\overline{w_2}w_3\overline{w_4}w_5\|_{0,b-1} \lesssim \prod_{j=1}^5 \|w_j\|_{s,1/2+}$$

in addition to similar estimates when D_x^s falls on one of the other functions. We shall prove only the displayed estimate, as the others are similar. We may estimate the $X^{0,b-1}$ norm by the $L_t^2 L_x^2$ norm. But then the claim follows from three applications of (3), two applications of (5), and Hölder (ensuring that the term with the D_x^s is estimated using (3)). □

6. Proof of Proposition 3.3. We can now prove Proposition 3.3, which as remarked before will give Theorem 1.1. Let T, w be as in the proposition. Our constants may depend on $\|w_0\|_2$ and $\|w_0\|_{H^s}$.

We start by rescaling the solution w. Let $\mu > 0$ be chosen later. We observe that w is a solution for the IVP (6) if and only if

$$w^\mu(t, x) = \frac{1}{\mu^{1/2}} w\left(\frac{t}{\mu^2}, \frac{x}{\mu}\right)$$

is a solution for the IVP (6) with initial data $w_0^\mu = \mu^{-1/2}w(\mu^{-1}x)$. From Plancherel's theorem and a simple computation we see that

$$\|I\partial_x w_0^\mu\|_2 \lesssim \frac{N^{1-s}}{\mu^s}\|w_0\|_{H^s},$$

while

$$\|Iw_0^\mu\|_2 \le \|w_0^\mu\|_2 = \|w_0\|_2 < \sqrt{2\pi}.$$

We now choose $\mu := N^{\frac{1-s}{s}}$. From the previous we see that $\|Iw_0^\mu\|_{H^1} \lesssim 1$, so from Sobolev embedding (or Gagliardo–Nirenberg) we obtain

$$E(Iw_0^\mu) \le C_1$$

for some constant $C_1 > 0$.

Now suppose inductively that we have a time T such that

$$E(Iw^\mu(T)) \le C_1 + C_2 N^{-1+} T,$$

where $C_2 > 0$ is a constant depending on C_1 to be chosen later. If $T \ll N^{1-}$, we then have $E(Iw^\mu(T)) \le 2C_1$, which implies from Lemma 3.6 that

$$\|Iw^\mu(T)\|_{H^1} \le C_3,$$

where C_3 depends on C_1. By Theorem 5.1 we thus have

$$\|Iw^\mu\|_{X^{1,1/2+}([T,T+\delta]\times\mathbb{R})} \le C_4,$$

where C_4, δ depend on C_3.

In the next four sections we shall prove the following key estimate.

Lemma 6.1. *For any Schwartz function w, we have*

$$\left| \int_T^{T+\delta} \Lambda_n(M_n; w(t))\, dt \right| \lesssim N^{-1+} \|Iw\|^n_{X^{1,1/2+}([T,T+\delta]\times\mathbb{R})} \tag{20}$$

for $n = 4, 6, 8$, where M_4, M_6, M_8 are as defined in Proposition 4.1.

Assuming this estimate for the moment, we see from the previous and Proposition 4.1 that

$$E(Iw^\mu(T+\delta)) \le E(Iw^\mu(T)) + C_5 N^{-1+},$$

where C_5 depends on δ and C_4. This allows us to close the induction hypothesis by setting $C_2 := C_5$. As a consequence, we have thus shown that[1]

$$\|Iw^\mu(T)\|_{H^1} \lesssim 1$$

for all $T \ll N^{1-}$. From the definition of I this implies that

$$\|w^\mu(T)\|_{H^s} \lesssim C_N$$

for all $T \ll N^{1-}$. Undoing the scaling, this implies that

$$\|w(T)\|_{H^s} \lesssim C_{N,\mu}$$

for all $T \ll N^{1-}/\mu^2$. However, if $s > 2/3$, then $N^{1-}/\mu^2 = N^{\frac{3s-2-}{s}}$ goes to infinity as $N \to \infty$, and Proposition 3.3 follows.

Remark 6.2. An examination of the above argument shows also that the H^s norm of w (and of u) grows at most polynomially in time; however, the order of the growth obtained by this argument goes to infinity as $s \to 2/3$.

[1] Strictly speaking, we have shown this only for T being an integer multiple of δ; however, this can be easily remedied, e.g., by using the fact that the $X^{1,1/2+}$ norm controls the $L_t^\infty H_x^1$ norm on $[T, T+\delta] \times \mathbb{R}$.

7. Proof of Lemma 6.1: Preliminaries. To prove Lemma 6.1 we shall treat the cases $n = 4$, $n = 6$, and $n = 8$ separately. The idea will be first to obtain some good estimates on M_n in terms of $m(\xi_i)$ and $\langle \xi_i \rangle$ and then to bound the resulting multilinear expression using standard tools such as the Strichartz estimates (3), (5), the trivial estimate

$$\|u\|_{L_t^2 L_x^2} \lesssim \|u\|_{0,0}, \tag{21}$$

and Hölder's inequality. In addition to the above linear estimates, we shall also take advantage of the following bilinear improvement to Strichartz's estimate in the case of differing frequencies (cf. [3, 21]).

Lemma 7.1. *For any Schwartz functions u, v with Fourier support in $|\xi| \sim R$, $|\xi| \ll R$, respectively, we have that*

$$\|uv\|_{L_t^2 L_x^2} = \|u\bar{v}\|_{L_t^2 L_x^2} \lesssim R^{-1/2} \|u\|_{0,1/2+} \|v\|_{0,1/2+}.$$

Proof. This is an improved Strichartz estimate of the type considered in [3]. In fact, the desired estimate is contained in Theorem 2 of [21]. We present the short proof for the sake of completeness.

It is enough to show that if u and v are solutions of the free Schrödinger equation, that is, $u = e^{it\partial_x^2}\phi$ and $v = e^{it\partial_x^2}\psi$, then

$$\|D_x^{1/2}(uv)\|_{L^2} \lesssim \|\phi\|_{L^2} \|\psi\|_{L^2}, \tag{22}$$

where D_x is the operator such that $\widehat{D_x f}(\xi) = \langle \xi \rangle \hat{f}(\xi)$. If we use duality and the change of variable $\xi_1 + \xi_2 = s$ and $|\xi_1|^2 + |\xi_2|^2 = r$, the left-hand side of (22) becomes

$$\begin{aligned} \sup_{\|F\|_{L^2} \le 1} &\int R^{1/2} F(\xi_1 + \xi_2, |\xi_1|^2 + |\xi_2|^2) \hat{\phi}(\xi_1) \hat{\psi}(\xi_2) d\xi_1 d\xi_2 \\ &\lesssim \int R^{1/2} F(s, r) \frac{H(s, r)}{R} ds dr, \end{aligned}$$

where $H(s, r)$ denotes the product of $\hat{\phi}$ and $\hat{\psi}$ in the new variables. Notice that the change of variables introduced above has a Jacobian of size R. Now if we use Cauchy–Schwarz and if we change the variables back to ξ_1 and ξ_2, then we obtain (22). □

In one of our subcases, we shall also take advantage of a trick (originally due to Bourgain [2]) of splitting the symbol $|\xi_1|^2 - \cdots + |\xi_n|^2$ as a sum of $\tau_j \mp |\xi_j|^2$.

Our estimates are not the best possible, and it is likely that one can improve the N^{-1+} gain in our estimates, probably to $N^{-3/2+}$. However, this will fall short of the N^{-2+} gain needed to push the global well-posedness down to match the local well-posedness theory at $s > 1/2$. However, one can recover this by adding higher-order correction terms to the energy $E_N(w(t))$, as in [8]. If one does this, one will end up estimating the Λ_6 and Λ_8 expressions rather than Λ_4. This will be beneficial because such expressions will have fewer derivatives in their symbol and can therefore enjoy better decay in N. The details of this argument will appear in a later paper.

We set out some notation. Let $n = 4$, 6, or 8, and let $\xi_1, \ldots, \xi_n$ be frequencies such that $\xi_1 + \cdots + \xi_n = 0$. Define $N_i := |\xi_i|$ and $N_{ij} := |\xi_{ij}|$. We adopt the notation that

$$1 \le \mathit{soprano}, \mathit{alto}, \mathit{tenor}, \mathit{baritone} \le n$$

are the distinct indices such that

$$N_{soprano} \geq N_{alto} \geq N_{tenor} \geq N_{baritone}$$

are the highest, second highest, third highest, and fourth highest values of the frequencies $N_1, \ldots, N_n$, respectively. (If there is a tie in frequencies, we break the tie arbitrarily.)

Since $\xi_1 + \cdots + \xi_n = 0$, we must have $N_{soprano} \sim N_{alto}$. Also, from Proposition 4.1 we see that M_n vanishes unless $N_{soprano} \gtrsim N$.

8. Proof of Lemma 6.1 when $n = 4$. We now estimate the Λ_4 expression. We begin by estimating the multiplier M_4.

LEMMA 8.1. *Let $\xi_1, \xi_2, \xi_3, \xi_4$ be such that $\xi_{1234} = 0$.*

- *If $N_{tenor} \sim N_{soprano}$, then*

$$|M_4(\xi_1, \xi_2, \xi_3, \xi_4)| \lesssim N^{-1}(N/N_{soprano})^{1/10}\langle \xi_{12}\xi_{14}\rangle^{1/2} \prod_{j=1}^{4} \langle \xi_j \rangle m(\xi_j). \tag{23}$$

- *If $N_{tenor} \ll N_{soprano}$, then*

$$|M_4(\xi_1, \xi_2, \xi_3, \xi_4)| \lesssim N^{-1}(N/N_{soprano})^{1/10} N_{soprano} \prod_{j=1}^{4} \langle \xi_j \rangle m(\xi_j). \tag{24}$$

Proof. Fix $\xi_1, \ldots, \xi_4$. If $N_{soprano} \ll N$, then M_4 vanishes by the second part of Proposition 4.1, and so we will assume that $N_{soprano} \gtrsim N$.

We split $M_4 = C_1 M_4' + C_2 M_4''$, where

$$M_4' := m_1 m_2 m_3 m_4 \xi_{12}\xi_{13}\xi_{14}$$

and

$$M_4'' := m_1^2 \xi_1^2 \xi_3 + m_2^2 \xi_2^2 \xi_4 + m_3^2 \xi_3^2 \xi_1 + m_4^2 \xi_4^2 \xi_2.$$

In the $N_{soprano} \gtrsim N$ case we will not need to exploit cancellation between M_4' and M_4'' (although such cancellation certainly exists) and shall estimate them separately.

Let us first prove (23). We begin with estimating M_4'. We have

$$\begin{aligned}
|M_4'| &= N_{12}N_{13}N_{14} m(N_1)m(N_2)m(N_3)m(N_4) \\
&\lesssim N_{12}N_{14}N_{soprano} m(N_{soprano})^3 \\
&\lesssim \langle N_{12}N_{14}\rangle^{1/2} N_{soprano}^2 m(N_{soprano})^3 \\
&\lesssim \langle N_{12}N_{14}\rangle^{1/2} \frac{1}{N}(N/N_{soprano})^{1/10} \langle N_{soprano}\rangle^3 m(N_{soprano})^3 \\
&\sim N^{-1}(N/N_{soprano})^{1/10}\langle N_{12}N_{14}\rangle^{1/2} \langle N_{soprano}\rangle m(N_{soprano}) \langle N_{alto}\rangle m(N_{alto}) \langle N_{tenor}\rangle m(N_{tenor}) \\
&\lesssim N^{-1}(N/N_{soprano})^{1/10}\langle N_{12}N_{14}\rangle^{1/2} \prod_{j=1}^{4} \langle N_j \rangle m(N_j),
\end{aligned}$$

as desired.

It remains to estimate M_4''. We divide it into two cases: $N_{baritone} \sim N_{soprano}$ and $N_{baritone} \ll N_{soprano}$.

Case 1. $N_{baritone} \sim N_{soprano}$.

In this case all the frequencies are comparable to each other. By symmetry we may assume that $N_{12} \leq N_{14}$, in which case it suffices to show

$$|M_4''| \lesssim N^{-1}(N/N_1)^{1/10} N_{12} N_1^4 m(N_1)^4.$$

We can rewrite $M_4'' = f(0) - f(h)$, where

$$f(h) := m(\xi_1 - h)^2(\xi_1 - h)^2(\xi_3 + h) + m(\xi_3 + h)^2(\xi_3 + h)^2(\xi_1 - h)$$

and $h := \xi_1 + \xi_2$. A routine calculation shows that

$$|f'(x)| \lesssim m(N_1)^2 N_1^2$$

for all $x = O(N_1)$, so by the mean value theorem and the assumption $N_1 \gtrsim N$ we have

$$|M_4''| = |f(0) - f(h)| \lesssim N_{12} m(N_1)^2 N_1^2 \lesssim N^{-1}(N/N_1)^{1/10} N_{12} N_1^4 m(N_1)^4,$$

as desired (in fact, we gain an additional power of N).

Case 2. $N_{baritone} \ll N_{soprano}$.

By symmetry we may assume that *baritone* $= 4$, and thus $N_1 \sim N_2 \sim N_3 \gg N_4$. In this case $N_{14} \sim N_1$, $N_{12} = N_{34} \sim N_1$, and $\langle N_4 \rangle m(N_4) \gtrsim 1$, so it suffices to show

$$|M_4''| \lesssim N^{-1}(N/N_1)^{1/10} N_1^4 m(N_1)^3.$$

But we may crudely estimate the left-hand side by

$$|M_4''| \lesssim m(N_1)^2 N_1^3 + m(N_4)^2 N_4^2 N_1 \lesssim m(N_1)^2 N_1^3,$$

which suffices since $N_1 \gtrsim N$. This proves (23).

Now we show (24). Observe that

$$N_{12} N_{13} N_{14} \lesssim N_{soprano}^2 N_{tenor},$$

and hence

$$\begin{aligned} |M_4'| &\lesssim N_{soprano}^2 N_{tenor} m(N_{soprano}) m(N_{alto}) m(N_{tenor}) m(N_{baritone}) \\ &\lesssim \prod_{j=1}^4 \langle N_j \rangle m(N_j) \\ &\lesssim N^{-1+} N_{soprano}^{1-} \prod_{j=1}^4 \langle N_j \rangle m(N_j). \end{aligned}$$

Thus it remains only to estimate M_4''. Since $\langle N_{baritone} \rangle m(N_{baritone})$ and $m(N_{tenor}) N^{-1}(N/N_{soprano})^{1/10} N_{soprano}$ are both $\gtrsim 1$, it suffices to show

$$|M_4''| \lesssim m(N_{soprano})^2 N_{soprano}^2 N_{tenor}.$$

By symmetry we may reduce to one of two cases.

Case 1. $N_3 = N_{tenor}$ and $N_4 = N_{baritone}$.

We crudely estimate

$$\begin{aligned}|M_4''(\xi_1,\xi_2,\xi_3,\xi_4)| &\lesssim m(N_1)^2 N_1^2 N_3 + m(N_2)^2 N_2^2 N_4 \\ &\quad + m(N_3)^2 N_3^2 N_1 + m(N_4)^2 N_4^2 N_2 \lesssim m(N_1)^2 N_1^2 N_3,\end{aligned}$$

as desired.

Case 2. $N_2 = N_{tenor}$ and $N_4 = N_{baritone}$.

In this case we estimate

$$\begin{aligned}|M_4''| &\lesssim |m_1^2\xi_1^2\xi_3 + m_3^2\xi_3^2\xi_1| + m(N_2)^2 N_2^2 N_4 + m(N_4)^2 N_4^2 N_2 \\ &= N_1 N_3 |m(\xi_1)^2\xi_1 - m(\xi_1+\xi_2+\xi_4)^2(\xi_1+\xi_2+\xi_4)| + O(m(N_1)^2 N_1^2 N_2).\end{aligned}$$

The function $m(\xi_1 + h)^2(\xi_1 + h)$ has a derivative of $O(m(N_1)^2)$ whenever $|h| \ll N_1$; thus by the mean value theorem we have

$$|M_4'(\xi_1,\xi_2,\xi_3,\xi_4)| \lesssim N_1 N_3 N_{24} m(N_1)^2 + O(m(N_1)^2 N_1^2 N_2) \lesssim m(N_1)^2 N_1^2 N_2,$$

as desired. □

We now prove (20) in the $n = 4$ case. It suffices to show that

$$\int_T^{T+\delta} \Lambda_4(M_4; w_1(t), \overline{w_2(t)}, w_3(t), \overline{w_4(t)})\, dt \lesssim N^{-1+} \prod_{j=1}^4 \|Iw_j\|_{1,1/2+}$$

for all Schwartz functions $w_1, \dots, w_4$ on $\mathbb{R} \times \mathbb{R}$. Since M_4 vanishes for $N_{soprano} \ll N$, it suffices by dyadic decomposition to show that

$$\begin{aligned}&\int_T^{T+\delta} \Lambda_4(M_4 \chi_{N_{soprano} \sim 2^k}; w_1(t), \overline{w_2(t)}, w_3(t), \overline{w_4(t)})\, dt \\ &\qquad \lesssim N^{-1+} 2^{(0+)k} (N/2^k)^{1/10} \prod_{j=1}^4 \|Iw_j\|_{1,1/2+}\end{aligned}$$

for all integers k for which $2^k \gtrsim N$. (The exact choice of the cutoff $\chi_{N_{soprano} \sim 2^k}$ is not important as we shall soon be taking absolute values everywhere anyway.)

Fix k. Without loss of generality we may assume that the Fourier transforms $\tilde{w}_j$ are real and nonnegative. We divide the Λ_4 integral into the regions $N_{tenor} \sim N_{soprano}$ and $N_{tenor} \ll N_{soprano}$.

Case 1. $N_{tenor} \sim N_{soprano}$.

We first perform some manipulations to eliminate the cutoff $\chi_{[T,T+\delta]}(t)$. Write $\chi_{[T,T+\delta]}(t) = a(t) + b(t)$, where $a(t)$ is $\chi_{[T,T+\delta]}(t)$ convolved with a smooth approximation to the identity of width 2^{-100k}, and $b(t) = \chi_{[T,T+\delta]}(t) - a(t)$.

Let us first consider the contribution of $b(t)$. We crudely estimate $M_4 = O(2^{10k})$ and estimate this contribution by

$$2^{10k} \int\int |b(t)||w_1(t,x)||w_2(t,x)||w_3(t,x)||w_4(t,x)|\, dxdt.$$

By Hölder, three applications of (3), one application of (4), and four applications of (15), we can bound this by

$$2^{10k} \|b\|_2 \prod_{j=1}^4 \|Iw_j\|_{1,1/2+}.$$

Since $\|b\|_2 \lesssim 2^{-50k}$, the claim then follows.

Now consider the contribution of $a(t)$. We use the following lemma.

LEMMA 8.2. *We have*

$$\|a(t)w_1\|_{1,1/2+} \lesssim 2^{(0+)k}\|w_1\|_{1,1/2+}.$$

Proof. By applying Plancherel, restricting to a single frequency ξ, and then undoing Plancherel, we see that it suffices to show that

$$\|a(t)f\|_{H_t^{1/2+}} \lesssim 2^{(0+)k}\|f\|_{H_t^{1/2+}}$$

for all functions f. However, this follows from the routine calculation

$$\|a(t)\|_{H_t^{1/2+}} \lesssim 2^{(0+)k}$$

and the fact that $H_t^{1/2+}$ is closed under multiplication. □

It therefore suffices to show

$$\left|\int \Lambda_4(M_4\chi_{N_{tenor}\sim N_{soprano}\sim 2^k}; w_1(t), \overline{w_2(t)}, w_3(t), \overline{w_4(t)})\ dt\right|$$
$$\lesssim N^{-1}(N/2^k)^{1/10}\prod_{j=1}^{4}\|Iw_j\|_{1,1/2+}.$$

Without loss of generality we may assume that the Fourier transforms $\tilde{w}_j$ are real and nonnegative. By Plancherel and (23) we estimate the left-hand side by

$$N^{-1}(N/2^k)^{1/10}\left|\int_* \langle\xi_{12}\xi_{14}\rangle^{1/2}\widetilde{ID_xw_1}(\tau_1,\xi_1)\widetilde{\overline{ID_xw_2}}(\tau_2,\xi_2)\widetilde{ID_xw_3}(\tau_3,\xi_3)\widetilde{\overline{ID_xw_4}}(\tau_4,\xi_4)\right|.$$

From the identity (cf. Bourgain [2] and Kenig–Ponce–Vega [17])

$$\begin{aligned}\sum_{j=1}^{4}(\tau_j-(-1)^{j-1}\xi_j^2) &= -\xi_1^2+\xi_2^2-\xi_3^2+\xi_4^2\\ &= \xi_{12}\xi_{2-1}+\xi_{34}\xi_{4-3}\\ &= \xi_{12}(\xi_{2-1}-\xi_{4-3})\\ &= -2\xi_{12}\xi_{14}\end{aligned}$$

we see that

$$\langle\xi_{12}\xi_{14}\rangle \lesssim \langle\tau_j-(-1)^{j-1}\xi_j^2\rangle$$

for some $j=1,2,3,4$. We shall assume $j=1$; the argument for other values of j is similar. We can then use duality and Plancherel to estimate the previous by

$$N^{-1}(N/2^k)^{1/10}\|Iw_1\|_{1,1/2+}\|\overline{ID_xw_2}ID_xw_3\overline{ID_xw_4}\|_{L_t^2L_x^2}.$$

However, this is acceptable by Hölder and three applications of (3).

Case 2. $N_{tenor} \ll N_{soprano}$.

We shall assume that *soprano* $=1$ and *alto* $=2$; the reader may verify that the other cases follow by the same argument. We may then restrict w_1, w_2 to have Fourier support in $|\xi|\sim 2^k$ and w_3, w_4 to have Fourier support in the region $|\xi|\ll 2^k$.

By (24) we have

$$|M_4| \lesssim N^{-1}(N/2^k)^{1/10}2^k\prod_{j=1}^{4}\langle\xi_i\rangle m(\xi_i).$$

The claim then follows from Hölder and two applications of Proposition 7.1.

9. Proof of Lemma 6.1 when $n = 6$. We begin with the analogue of Lemma 8.1.

LEMMA 9.1. *Let $\xi_1, \ldots, \xi_6$ be such that $\xi_{123456} = 0$.*

- *If $N_{tenor} \sim N_{soprano}$, then*

$$(25) \quad |M_6(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6)| \lesssim N^{-1} \langle \xi_{soprano} \rangle m(\xi_{soprano}) \langle \xi_{alto} \rangle m(\xi_{alto}) \langle \xi_{tenor} \rangle m(\xi_{tenor}).$$

- *If $N_{tenor} \ll N_{soprano}$, then*

$$(26) \quad |M_6(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6)| \lesssim N^{-1+} \langle N_{soprano} \rangle^{1-} \langle \xi_{soprano} \rangle m(\xi_{soprano}) \langle \xi_{alto} \rangle m(\xi_{alto}).$$

One can improve these estimates by exploiting further cancellation in the expression M_6, but we shall not need to do so because of the good smoothing properties of (6).

Proof. Since $\xi_{123456} = 0$, we have $N_{alto} \sim N_{soprano}$. We may also assume that $N_{soprano} \gtrsim N$ since M_6 vanishes otherwise.

We have the very crude estimate

$$|M_6| \lesssim N_{soprano}^2.$$

If $N_{tenor} \sim N_{soprano}$, we then have

$$|M_6| \lesssim N_{soprano}^2 \lesssim N^{-1} m(N_{soprano}) N_{soprano} m(N_{alto}) N_{alto} m(N_{tenor}) N_{tenor}$$

(using the hypothesis $s > 2/3$), and (25) follows.

Now suppose that $N_{tenor} \ll N_{soprano}$. Then

$$|M_6| \lesssim N_{soprano}^2 \lesssim N^{-1+} \langle N_{soprano} \rangle^{1-} m(N_{soprano}) N_{soprano} m(N_{alto}) N_{alto}$$

(since $s > 1/2$), and (26) follows. □

We now prove (20) for $n = 6$. As in the previous section, it suffices to show

$$\int_T^{T+\delta} \Lambda_6(M_6; w_1(t), \overline{w_2(t)}, w_3(t), \overline{w_4(t)}, w_5(t), \overline{w_6(t)})\ dxdt \lesssim N^{-1+} \prod_{j=1}^{6} \|Iw_j\|_{1,1/2-}$$

for all Schwartz functions $w_1, \ldots, w_6$ on $\mathbb{R} \times \mathbb{R}$. Without loss of generality we may assume that the Fourier transforms $\tilde{w}_i$ of w_i are real and nonnegative.

We again divide the analysis into the cases $N_{tenor} \sim N_{soprano}$ and $N_{tenor} \ll N_{soprano}$.

Case 1. $N_{tenor} \sim N_{soprano}$.

By (25) and symmetry it suffices to show

$$\int_T^{T+\delta} \int \prod_{j=1}^{3} |D_x I w_j| \prod_{j=4}^{6} |w_j|\ dxdt \lesssim \prod_{j=1}^{6} \|Iw_j\|_{1,1/2+}.$$

However, this follows from Hölder, six applications of (3) first, and three applications of (15) later.

Case 2. $N_{tenor} \ll N_{soprano}$.

We shall assume that *soprano* = 1 and *alto* = 2; the reader may verify that the other cases follow by the same argument.

First suppose that $N_{soprano} \sim 2^k$ for some integer k. Then w_1, w_2 have Fourier support on $|\xi| \sim 2^k$, while w_3, w_4, w_5, w_6 have Fourier support on $|\xi| \ll 2^k$.

We apply (26) and bound the contribution of this case by

$$N^{-1+}2^{(1-)k} \int_T^{T+\delta} \int \prod_{j=1}^{2} |D_x I w_j| \prod_{j=3}^{6} |w_j| \; dxdt,$$

which we bound using Hölder by

$$N^{-1+}2^{(1-)k} \|(D_x I w_1) w_3\|_{L_t^2 L_x^2} \|(D_x I w_2) w_4\|_{L_t^2 L_x^2} \|w_5\|_{L_t^\infty L_x^\infty} \|w_6\|_{L_t^\infty L_x^\infty}.$$

By Lemma 7.1, (5), and (15) we can bound this by

$$N^{-1+}2^{(0-)k} \prod_{j=1}^{6} \|I w_j\|_{1,1/2+}.$$

The claim then follows by summing in k.

10. Proof of Lemma 6.1 when n = 8. We begin with the analogue of Lemma 8.1.

LEMMA 10.1. *For any* $\xi_1, \dots, \xi_6$ *with* $\xi_{123456} = 0$*, we have*

$$|M_8(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6, \xi_7, \xi_8)| \lesssim N^{-1} \langle \xi_{soprano} \rangle m(\xi_{soprano}) \langle \xi_{alto} \rangle m(\xi_{alto}). \tag{27}$$

Proof. As usual, we may assume that $N_{soprano} \sim N_{alto} \gtrsim N$. We crudely estimate

$$|M_8| \lesssim N_{soprano} \lesssim N^{-1} N_{soprano} m(N_{soprano}) N_{alto} m(N_{alto}),$$

and the claim follows. □

To prove (20) for $n = 8$ it suffices to show

$$\int_T^{T+\delta} \Lambda_8(M_8; w_1(t), \dots, \overline{w_8(t)}) \; dt \lesssim N^{-1} \prod_{j=1}^{8} \|I w_j\|_{1,1/2+}$$

for all Schwartz functions $w_1, \dots, w_8$ on $\mathbb{R} \times \mathbb{R}$. Without loss of generality, we may assume that the Fourier transforms $\tilde{w_i}$ of w_i are real and nonnegative. By Lemma 10.1 and symmetry it thus suffices to show

$$\int_T^{T+\delta} \int |D_x I w_1| |D_x I w_2| \prod_{j=3}^{8} |w_j| \; dxdt \lesssim N^{-1} \prod_{j=1}^{8} \|I w_j\|_{1,1/2+}.$$

However, this follows from Hölder, six applications of (3), and two applications of (5) and (15).

Remark 10.2. As was shown in section 3, the gauge transform in Definition 3.1 introduces a quintic term in the IVP (6). Then one can ask if the same arguments we proposed above can be used in order to study the global well-posedness of the quintic nonlinear Schrödinger IVP

$$\begin{cases} i\partial_t v + \partial_x^2 v + \lambda |v|^4 v = 0, & \\ v(x,0) = v_0(x), & x \in \mathbb{R}, \quad t \in \mathbb{R}, \end{cases} \tag{28}$$

where $\lambda \in \mathbb{R}$. In this case we define the energy

$$H(f) := \int |\partial_x f(x)|^2 \, dx - \frac{\lambda}{6} \int |f|^2 \, dx.$$

By Plancherel, we may write $H(f)$ using the Λ notation as

$$H(f) = \Lambda_2(\xi_1\xi_2; f) - \frac{\lambda}{6}\Lambda_6(1; f).$$

As in Lemma 3.5, one can prove that the energy $H(v(t))$ of the solution v for (28) is constant. Now let us define the new energy

$$H_N(v) = H(Iv) = \Lambda_2(\xi_1\xi_2 m_1 m_2; v) - \frac{\lambda}{6}\Lambda_6\left(\prod_{i=1}^{6} m_i; v\right)$$

just like we did in section 4. Then, by the analogue of (8), $\partial_t H_N(v(t))$ will involve terms of type Λ_2, Λ_6, and Λ_{10}. Using the same ideas presented in the proof of Lemma 6.1, we can also estimate in the appropriate way the term involving Λ_{10}. If in (28) we assume that $\lambda < 0$ (defocusing) or that the L^2 norm of the initial data is small (so that the Gagliardo–Niremberg inequality can be applied), then the energy $H(v)(t)$ stays positive for all times, and global well-posedness in H^s for $s > 2/3$ will follow. We will present the details of the proof in a future paper. It has to be said here that global results for "small data" are already available for (28) through more standard arguments [6].

REFERENCES

[1] H. Biagioni and F. Linares, *Ill-posedness for the derivative Schrödinger and generalized Benjamin–Ono equations*, Trans. Amer. Math. Soc., 353 (2001), pp. 3649–3659.

[2] J. Bourgain, *Fourier restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations*, I, Geom. Funct. Anal., 3 (1993), pp. 107–156.

[3] J. Bourgain, *Refinements of Strichartz' inequality and applications to 2D-NLS with critical nonlinearity*, Internat. Math. Res. Notices, 5 (1998), pp. 253–283.

[4] J. Bourgain, *Periodic Korteweg-de Vries equation with measures as initial data*, Selecta Math. (N.S.), 3 (1997), pp. 115–159.

[5] J. Bourgain, *Global Solutions of Nonlinear Schrödinger Equations*, Amer. Math. Soc. Colloq. Publ. 46, AMS, Providence, RI, 1999.

[6] T. Cazenave and F. Weissler, *The Cauchy problem for the critical nonlinear Schrödinger equation in H^s*, Nonlinear Anal., 14 (1990), pp. 807–836.

[7] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Global Well-Posedness in $H^{4/7+}$ for the 2D Cubic NLS*, in preparation.

[8] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Sharp Global Well-Posedness for Periodic and Non-Periodic KdV and mKdV*, in preparation.

[9] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Multilinear Estimates for Periodic KdV Equations, and Applications*, in preparation.

[10] N. Hayashi, *The initial value problem for the derivative nonlinear Schrödinger equation in the energy space*, Nonlinear Anal., 20 (1993), pp. 823–833.

[11] N. Hayashi and T. Ozawa, *On the derivative nonlinear Schrödinger equation*, Phys. D, 55 (1992), pp. 14–36.

[12] N. Hayashi and T. Ozawa, *Finite energy solutions of nonlinear Schrödinger equations of derivative type*, SIAM J. Math. Anal., 25 (1994), pp. 1488–1503.

[13] N. Hayashi and T. Ozawa, *Remarks on nonlinear Schrödinger equations in one space dimension*, Differential Integral Equations, 2 (1994), pp. 453–461.

[14] M. Keel and T. Tao, *Local and global well-posedness of wave maps on $\mathbb{R}^{1+1}$ for rough data*, Internat. Math. Res. Notices, 21 (1998), pp. 1117–1156.

[15] M. Keel and T. Tao, *Global Well-Posedness of the Maxwell-Klein-Gordon Equation below the Energy Norm*, in preparation.

[16] C. E. Kenig, G. Ponce, and L. Vega, *The Cauchy problem for the Korteweg-de Vries equation in Sobolev spaces of negative indices*, Duke Math J., 71 (1993), pp. 1–21.

[17] C. Kenig, G. Ponce, and L. Vega, *A bilinear estimate with applications to the KdV equation*, J. Amer. Math. Soc., 9 (1996), pp. 573–603.

[18] W. Mio, T. Ogino, K. Minami, and S. Takeda, *Modified nonlinear Schrödinger for Alfvén waves propagating along the magnetic field in cold plasma*, J. Phys. Soc. Japan, 41 (1976), pp. 265–271.

[19] E. Mjolhus, *On the modulational instability of hydromagnetic waves parallel to the magnetic field*, J. Plasma Phys., 16 (1996), pp. 321–334.

[20] T. Ozawa, *On the nonlinear Schrödinger equations of derivative type*, Indiana Univ. Math. J., 45 (1996), pp. 137–163.

[21] T. Ozawa and Y. Tsutsumi, *Space-time estimates for null gauge forms and nonlinear Schrödinger equations*, Differential Integral Equations, 11 (1998), pp. 201–222.

[22] H. Takaoka, *Well-posedness for the one dimensional Schrödinger equation with the derivative nonlinearity*, Adv. Differential Equations, 4 (1999), pp. 561–680.

[23] C. Sulem and P.-L. Sulem, *The Nonlinear Schrödinger Equation. Self-Focusing and Wave Collapse*, Appl. Math. Sci. 139, Springer-Verlag, New York, 1999.

[24] H. Takaoka, *Global well-posedness for Schrödinger equations with derivative in a nonlinear term and data in low-order Sobolev spaces*, Electron. J. Differential Equations, 43 (2001).

[25] T. Tao, *Multilinear weighted convolution of L^2 functions and applications to nonlinear dispersive equations*, Amer. J. Math., to appear.

[26] M. Tsutsumi and I. Fukuda, *On solutions of the derivative nonlinear Schrödinger equation: Existence and uniqueness theorem*, Funkcial. Ekvac., 23 (1980), pp. 259–277.

[27] M. Tsutsumi and I. Fukuda, *On solutions of the derivative nonlinear Schrödinger equation* II, Funkcial. Ekvac., 24 (1981), pp. 85–94.

[28] M. I. Weinstein, *Nonlinear Schrödinger equations and sharp interpolation estimates*, Comm. Math. Phys., 87 (1983), pp. 567–576.

SIAM J. MATH. ANAL.
Vol. 33, No. 3, pp. 670–685

NEW STABILITY ESTIMATES FOR THE INVERSE ACOUSTIC INHOMOGENEOUS MEDIUM PROBLEM AND APPLICATIONS[*]

PETER HÄHNER[†] AND THORSTEN HOHAGE[‡]

Abstract. This paper is concerned with the scattering of time-harmonic acoustic waves by inhomogeneous media. We study the problem to recover the refractive index from far field measurements and from near field mesurements. We establish logarithmic stability estimates for these problems using a priori information with respect to Sobolev norms and a priori information about the support of the inhomogeneity. Our results improve previous estimates due to Stefanov by giving an explicit exponent in the logarithmic estimate, by using the L^2-norm for far field patterns, and by dropping the assumption that the refractive indices are close together. These improvements make it possible to prove convergence rates for iterative regularization methods.

Key words. inverse medium scattering, logarithmic stability, regularization methods

AMS subject classifications. 35R30, 81U40, 35J05, 65J20

PII. S0036141001383564

1. Introduction.

The scattering of time-harmonic acoustic waves in an inhomogeneous medium with refractive index n is described by the differential equation

$$\Delta u + k^2 n u = 0. \tag{1.1}$$

Here the real part of the complex valued function u describes the space-dependent part of a velocity potential, and $k > 0$ is the wave number. We assume that the refractive index n is a complex valued function in $\mathbb{R}^3$ satisfying $\mathrm{Im}(n(x)) \geq 0$ for all $x \in \mathbb{R}^3$ and supp $(1 - n) \subset B_1$. (For $R > 0$ we use the notation $B_R := \{x \in \mathbb{R}^3 \colon |x| < R\}$.) We will also assume throughout that $1 - n$ belongs to the Sobolev space $H^s(\mathbb{R}^3)$ for some fixed $s > 3/2$. Due to Sobolev's imbedding theorem, this implies that n is uniformly Hölder continuous in $\mathbb{R}^3$ with Hölder exponent $s - 3/2$. Since by rescaling we can always obtain that $\mathrm{supp}(1 - n) \subset B_1$, this is not a severe restriction.

The following problem will be referred to as the *direct scattering problem*: Given an incident wave $u^{\mathrm{i}} \in C^2(B_1)$ satisfying the Helmholtz equation $\Delta u^{\mathrm{i}} + k^2 u^{\mathrm{i}} = 0$ in B_1 and k and n as above, find a scattered wave $u^{\mathrm{s}} \in C^2(\mathbb{R}^3)$ satisfying the differential equation

$$\Delta u^{\mathrm{s}} + k^2 n u^{\mathrm{s}} = k^2 (1 - n) u^{\mathrm{i}} \quad \text{in } \mathbb{R}^3 \tag{1.2a}$$

and the Sommerfeld radiation condition

$$\lim_{|x| \to \infty} |x| \left(\frac{\partial u^{\mathrm{s}}}{\partial r}(x) - iku^{\mathrm{s}}(x) \right) = 0 \tag{1.2b}$$

[*] Received by the editors January 15, 2001; accepted for publication (in revised form) May 16, 2001; published electronically October 4, 2001.
http://www.siam.org/journals/sima/33-3/38356.html

[†] Inst. f. Num. u. Angew. Math., Universität Göttingen, Lotzestraße 16–18, D-37083 Göttingen, Germany (haehner@math.uni-goettingen.de). Parts of this article are taken from this author's habilitation thesis [7].

[‡] Konrad-Zuse Zentrum, Takustr. 7, D-14195 Berlin, Germany (hohage@zib.de). Part of this author's work was carried out at the University of Linz, Austria, with support of the Austrian Science Fund (FWF) under the project SFB F013 "Numerical and Symbolic Scientific Computing."

uniformly for all directions $\hat{x} = (1/r)x$, where $r := |x|$. The right-hand side of (1.2a) is defined to be zero outside of B_1. It can be shown that this problem has a unique solution (cf. [3, Chapter 8]). Note that the total field $u := u^{\mathrm{i}} + u^{\mathrm{s}}$ satisfies (1.1).

To formulate our first inverse problem, let us introduce point sources $u^{\mathrm{i}}(x) = \Phi(x, y)$, located at $y \in \mathbb{R}^3$, $|y| > 1$, as incident fields. Here

$$\Phi(x, y) := \frac{1}{4\pi} \frac{e^{ik|x-y|}}{|x-y|}, \quad x, y \in \mathbb{R}^3, \quad x \neq y,$$

denotes the fundamental solution to the Helmholtz equation. The corresponding scattered fields u^{s} satisfying (1.2a) and (1.2b) for $u^{\mathrm{i}}(x) = \Phi(x, y)$ are denoted by $w_n^{\mathrm{s}}(x, y)$. The total field $w_n(x, y) := \Phi(x, y) + w_n^{\mathrm{s}}(x, y)$ is the Green function for our scattering problem. It is actually defined for all x, $y \in \mathbb{R}^3$, $x \neq y$, but we shall need it only for points $|y| > 1$. We consider the inverse problem to determine the refractive index n from a knowledge of its corresponding Green function w_n on $\partial B_R \times \partial B_R$ with some $R > 1$. This means we are probing the medium using point sources located on ∂B_R and measure the scattered waves on ∂B_R in order to recover the inhomogeneity. We will show the following logarithmic stability estimate.

Theorem 1.1. *Let $C_n > 0$ and $R > 1$ be given constants. Then, there exists a positive constant C (depending only on s, k, R, and C_n) such that for all refractive indices n, $\tilde{n}$ satisfying $\|1 - n\|_{H^s}, \|1 - \tilde{n}\|_{H^s} \leq C_n$, $\operatorname{supp}(1 - n)$, $\operatorname{supp}(1 - \tilde{n}) \subset B_1$, the estimates*

$$\|n - \tilde{n}\|_{L^2(B_1)} \leq C \left[-\ln^- \left(\|w_n^{\mathrm{s}} - w_{\tilde{n}}^{\mathrm{s}}\|_{L^2(\partial B_R \times \partial B_R)} \right) \right]^{-\frac{s}{s+3}} \tag{1.3}$$

and

$$\|n - \tilde{n}\|_{\infty} \leq C \left[-\ln^- \left(\|w_n^{\mathrm{s}} - w_{\tilde{n}}^{\mathrm{s}}\|_{L^2(\partial B_R \times \partial B_R)} \right) \right]^{-\frac{2s-3}{2s+3}} \tag{1.4}$$

hold true. Here $\ln^-(t) := \ln(t)$ for $t \leq \exp(-1)$, and $\ln^-(t) := -1$ else.

This theorem implies that if one uses point sources located on ∂B_R to probe the medium and if one measures the scattered waves on ∂B_R, then the refractive index depends continuously on these data under the a priori assumptions $\|1 - n\|_{H^s} \leq C_n$ and $\operatorname{supp}(1 - n) \subset B_1$. The logarithmic estimates (1.3) and (1.4) reflect the exponentially ill-posed nature of the problem. Let us point out that Theorem 1.1 also implies a global uniqueness result for the inverse scattering problem to determine the refractive index from measurements of the scattered fields on ∂B_R using point sources located on ∂B_R. Namely, two refractive indices from H^s having the same Green function on $\partial B_R \times \partial B_R$ must be identical.

Another interesting inverse problem results from using far field data instead of near field data. A solution v to the Helmholtz equation in an exterior domain satisfying (1.2b) is known to obey

$$v(x) = \frac{e^{ik|x|}}{|x|} \left(v^{\infty}\left(\frac{x}{|x|} \right) + \mathrm{O}\left(\frac{1}{|x|} \right) \right)$$

(cf. [3, Section 2.2]). The function $v^{\infty} : S^2 \to \mathbb{C}$, $S^2 := \{x \in \mathbb{R}^3 : |x| = 1\}$ is called *the far field pattern* of v. In this context we will consider plane incident waves $u^{\mathrm{i}}(x) = u^{\mathrm{i}}(x, d) = \exp(ikx \cdot d)$, $d \in S^2$. The corresponding scattered fields are denoted by $u_n^{\mathrm{s}}(\cdot, d)$, and their far field patterns are denoted by $u_n^{\infty}(\cdot, d)$. Probing the medium with plane incident waves from all directions and measuring the corresponding far field patterns lead to the problem to recover n from a knowledge of u_n^∞ on $S^2 \times S^2$. We will prove the following result.

THEOREM 1.2. *Let $C_n > 0$ and $0 < \epsilon < \frac{s}{s+3}$ be given constants. Then there exists a positive constant C (depending only on s, ϵ, k, and C_n) such that for all refractive indices n, $\tilde{n}$ satisfying $\|1-n\|_{H^s}, \|1-\tilde{n}\|_{H^s} \le C_n$, and $\operatorname{supp}(1-n)$, $\operatorname{supp}(1-\tilde{n}) \subset B_1$, the estimate*

$$\|n - \tilde{n}\|_{L^2(B_1)} \le C \left[-\ln^- \left(\|u_n^\infty - u_{\tilde{n}}^\infty\|_{L^2(S^2\times S^2)}\right)\right]^{-\frac{s}{s+3}+\epsilon} \tag{1.5}$$

holds true.

Moreover, for $0 < \epsilon < \frac{2s-3}{2s+3}$ the maximum norm of $n - \tilde{n}$ can be estimated by

$$\|n - \tilde{n}\|_\infty \le C \left[-\ln^- \left(\|u_n^\infty - u_{\tilde{n}}^\infty\|_{L^2(S^2\times S^2)}\right)\right]^{-\frac{2s-3}{2s+3}+\epsilon}. \tag{1.6}$$

The main idea for the proof of Theorem 1.1 goes back to Alessandrini [1], who examined the continuous dependence of the conductivity on its corresponding Dirichlet-to-Neumann map. In [17] Stefanov applied this idea to potential scattering problems and proved that the potential depends continuously on the corresponding Dirichlet-to-Neumann map. Both articles employ a lemma about holomorphic functions in several variables which prevents the exponent in the inequalities from being known explicitly. In order to obtain the explicit exponents $-s/(s+3)$ and $-(2s-3)/(2s+3)$, we present a modified and simpler proof. Moreover, contrary to Stefanov, who establishes his result only for refractive indices n and $\tilde{n}$ sufficiently close to each other, we shall need only the assumption that the H^s norms of $1-n$ and $1-\tilde{n}$ are bounded by a constant C_n. Stefanov's paper [17] also contains a stability estimate involving far field data. However, instead of the L^2 norm, a very strong norm involving exponentially increasing weights for the Fourier coefficients of the far field patterns is used. Since no estimate on the measurement error with respect to this norm will be available in practice, Stefanov's result is primarily of theoretical interest.

Let us point out that the function $[-\ln t]^{-1}$ is converging very slowly to 0 as $t \to 0$. The power $0 < p < 1$ in the function $[-\ln t]^{-p}$ is even deteriorating the convergence. However, since it is possible to prove convergence rates for iterative regularization methods which are based on stability estimates and since our numerical example indicates that the convergence of an iterative regularization method is not much faster than predicted by our estimate, we have strong evidence that our stability estimate cannot be significantly improved.

We also want to emphasize that assuming high regularity for n, i.e., using the a priori information $1 - n \in H_0^s(B_1)$ with a very large s, will improve the estimate somewhat but not very much.

In the next section we briefly collect the necessary results for the Green function. The following two sections are dedicated to the proofs of the two main theorems. In the final section of this paper we show how Theorem 1.2 can be used to obtain convergence rates for iterative regularization methods, and we present numerical computations comparing the actual speed of convergence to our estimates.

2. The Green function.

In this section we want to briefly review the properties of the Green function that we shall use in what follows for our stability estimates.

Let us first introduce the *Lippmann–Schwinger equation*, which is an equivalent formulation of (1.2a), (1.2b) as an integral equation of the second kind:

$$u(x) + k^2 \int_{B_1} \Phi(x,y)\,(1-n(y))\,u(y)\,\mathrm{d}y = u^{\mathrm{i}}(x), \quad x \in B_1. \tag{2.1}$$

If $u \in C(B_1)$ is a solution to (2.1), then

$$u^{\mathrm{s}}(x) := -k^2 \int_{B_1} \Phi(x,y)\,(1-n(y))\,u(y)\,\mathrm{d}y, \quad x \in \mathbb{R}^3, \tag{2.2}$$

is the solution to (1.2a), (1.2b), and vice versa (cf. [3, Section 8.2]). Using the incident waves $u^{\mathrm{i}}(x) = \Phi(x,y)$, we find that w_n^s is infinitely smooth for $|x|, |y| > 1$, $x \neq y$. From (2.2) and the asymptotic behavior of $\Phi(x,y)$ we obtain

$$u^\infty(\hat{x}) = -\frac{k^2}{4\pi} \int_{B_1} e^{-ik\hat{x}\cdot y}\,(1-n(y))\,u(y)\,\mathrm{d}y, \quad \hat{x} \in S^2.$$

Hence the far field u^∞ is C^∞ on $S^2 \times S^2$.

For a fixed radius $R > 1$ let us introduce the single layer potential having the Green function $w_n(x,y) = \Phi(x,y) + w_n^s(x,y)$ as kernel, that is, the function

$$u(x) := \int_{\partial B_R} w_n(x,y)\,\varphi(y)\;\mathrm{d}s(y), \quad x \in \mathbb{R}^3. \tag{2.3}$$

Here we assume that the density φ is a continuous function on ∂B_R. For convenience we also introduce the operator S_n which maps φ to $u|_{\partial B_R}$, i.e.,

$$(S_n\varphi)(x) := \int_{\partial B_R} w_n(x,y)\,\varphi(y)\;\mathrm{d}s(y), \quad x \in \partial B_R. \tag{2.4}$$

Lemma 2.1. *Suppose $1 < R$. Then*

(a) *The Green function satisfies $w_n(x,y) = w_n(y,x)$. In particular, the operator S_n is symmetric: $\int_{\partial B_R} g\,(S_n f)\;\mathrm{d}s = \int_{\partial B_R} (S_n g) f\;\mathrm{d}s$ for all f, $g \in C(\partial B_R)$.*

(b) *The mapping $1 - n \mapsto w_n|_{\partial B_R \times \partial B_R}$, regarded as a mapping from $H_0^s(B_1)$ to $L^2(\partial B_R \times \partial B_R)$, is completely continuous. In particular, bounded sets are mapped into bounded sets.*

Proof. The symmetry of w_n^{s} and then of the Green function w_n can be shown with the help of Green's theorem (cf. [16, Section 8, Chapter 1.2.2 and Section 3, Chapter 1.2.3]). This immediately implies the symmetry of S_n by interchanging the order of integration.

For part (b) we observe that the embedding $H_0^s(B_1) \to H_0^{s'}(B_1)$ with $s > s' > 3/2$ is a compact linear mapping. Since the mapping $H_0^{s'}(B_1) \to L^2(\partial B_R \times \partial B_R)$, $1 - n \mapsto w_n|_{\partial B_R \times \partial B_R}$ is continuous, the composition is completely continuous and maps bounded sets into bounded sets. □

The next lemma states that the single layer potential having as kernel the Green function w_n is the unique solution to a certain transmission problem.

Lemma 2.2. *For $f \in C(\partial B_R)$ the function*

$$u(x) := \int_{\partial B_R} w_n(x,y)\,f(y)\;\mathrm{d}s(y), \quad x \in \mathbb{R}^3, \tag{2.5}$$

is the unique solution to the boundary value problem

$$\Delta u + k^2 n u = 0 \quad \text{in } \mathbb{R}^3 \setminus \partial B_R, \tag{2.6a}$$

$$\frac{\partial u_-}{\partial \nu} - \frac{\partial u_+}{\partial \nu} = f \quad \text{on } \partial B_R, \tag{2.6b}$$

$$\lim_{r\to\infty} r\left(\frac{\partial u}{\partial r} - iku\right) = 0 \tag{2.6c}$$

in $C(\mathbb{R}^3)\cap C^2(\mathbb{R}^3\backslash\partial B_R)$*. Here,* ν *denotes the unit normal vector on* ∂B_R *directed into the exterior of* B_R*.*

Proof. In order to prove that the boundary value problem (2.6) has at most one solution, we assume that u is a solution of (2.6) with $f=0$. Then we can follow the first part of the reasoning for the uniqueness proof of (1.2a), (1.2b) [3, Theorem 8.7], and we obtain $u=0$ in the exterior of B_R, whence $u=\frac{\partial u_-}{\partial\nu}=0$ on ∂B_R. Now Green's representation formula (cf. [3, Theorem 2.1]),

$$u(x)=\int_{\partial B_R}\left\{\frac{\partial u}{\partial\nu}(y)\,\Phi(x,y)-u(y)\,\frac{\partial\Phi(x,y)}{\partial\nu(y)}(x,y)\right\}\,\mathrm{d}s(y)$$
$$-\int_{B_R}(\Delta u(y)+k^2u(y))\,\Phi(x,y)\,\mathrm{d}y,\quad x\in B_R,$$

together with the differential equation (2.6a), implies that u is a solution of the homogeneous Lippmann–Schwinger equation. Thus we also have $u=0$ in B_R.

Since u defined as in (2.5) is a superposition of the solutions $w_n(\cdot,y)$ to $\Delta_x w_n+k^2nw_n=0$, u itself also satisfies this differential equation in $\mathbb{R}^3\setminus\partial B_R$. Moreover, u is a radiating solution. Finally, we can conclude from the regularity properties of the single layer potential with kernel Φ (cf. [3, Section 3.1]) that u as defined in (2.5) satisfies the boundary conditions. Hence it is the solution of (2.6). □

Our final lemma of this section states a series representation of w_n^s and of u_n^∞ in terms of spherical harmonics. For convenience we want to use the notation

$$\mathcal{M}:=\{(l_1,m_1,l_2,m_2)\in\mathbb{N}_0\times\mathbb{Z}\times\mathbb{N}_0\times\mathbb{Z}:|m_1|\le l_1\text{ and }|m_2|\le l_2\},$$

and Y_l^m, $l=0,1,2,\ldots$, $-l\le m\le l$, are a basis of the spherical harmonics. $h_l^{(1)}$ denotes the spherical Hankel of the first kind and of order l (cf., e.g., [3, Section 2.4]).

LEMMA 2.3.

(a) *For* $|x|,\ |y|>1$ *the scattered part of the Green function can be represented as*

$$w_n^s(x,y)=-\frac{k^2}{4\pi}\sum_{(l_1,m_1,l_2,m_2)\in\mathcal{M}} i^{l_1-l_2}\,\alpha_{l_1,m_1,l_2,m_2}\,h_{l_1}^{(1)}(k|x|)\,h_{l_2}^{(1)}(k|y|)$$
$$\times\,Y_{l_1}^{m_1}\left(\frac{x}{|x|}\right)Y_{l_2}^{m_2}\left(\frac{y}{|y|}\right)$$

with suitable constants α_{l_1,m_1,l_2,m_2}*. The series converges absolutely and uniformly on compact subsets of* $\mathbb{R}^3\setminus\overline{B_1}$ *and can be differentiated termwise.*

(b) *The far field* $u_n^\infty(\hat{x},d)$ *has the expansion*

$$u_n^\infty(\hat{x},d)=\sum_{(l_1,m_1,l_2,m_2)\in\mathcal{M}}\alpha_{l_1,m_1,l_2,m_2}\,Y_{l_1}^{m_1}(\hat{x})\,Y_{l_2}^{m_2}(d),$$

where the coefficients α_{l_1,m_1,l_2,m_2} *are the same as in part* (a)*.*

For a proof we refer to Stefanov [17, Proposition 2.2].

3. Estimating the refractive index with the help of near field data (proof of Theorem 1.1). The main idea for the stability estimate is to estimate the Fourier coefficients in the Fourier expansion of $n-\tilde{n}$ with the help of special solutions to (1.1) which depend in a particular way on a complex parameter vector $\zeta\in\mathbb{C}^3$. We call these solutions *geometrical optics solutions.*

LEMMA 3.1. *Suppose $k > 0$, $1 < R'$, a positive constant C_n, and a refractive index n with $\|1-n\|_{H^s} \leq C_n$ are given. Then there are constants $M_1, M_2 > 0$, depending only on s, k, R', and C_n with the following property. For all $\zeta \in \mathbb{C}^3$ satisfying*

$$|\operatorname{Im}(\zeta)| \geq M_1 \quad \textit{and} \quad \zeta \cdot \zeta = k^2$$

there exists a function $v(\cdot\,,\zeta) \in C^2(B_{R'})$ such that

$$U(x,\zeta) := e^{i\zeta\cdot x}(1+v(x,\zeta)), \quad x \in B_{R'}, \tag{3.1}$$

is a solution to $\Delta u + k^2 n u = 0$ in $B_{R'}$, and the estimate

$$\|v(\cdot\,,\zeta)\|_{L^2(B_{R'})} \leq \frac{M_2}{|\operatorname{Im}(\zeta)|} \tag{3.2}$$

holds.

These solutions have a long history and were employed by different authors dealing with inverse scattering problems. We refer the reader to [15] for a brief overview. A proof of the above lemma for the equation $\Delta u + qu = 0$ can be found in [18]. For our case we refer the reader to Lemma 5 in [6].

We have

$$n(x) - \tilde{n}(x) = \sum_{\gamma\in\mathbb{Z}^3} \widehat{(n-\tilde{n})}(\gamma)\, \frac{1}{(2\pi)^{3/2}}\, \exp(i\gamma\cdot x), \quad x \in B_1, \tag{3.3}$$

where $\hat{f}(\gamma)$ denotes the Fourier coefficients of a function $f \in L^2((-\pi,\pi)^3)$ with respect to the orthonormal bases $(2\pi)^{-3/2}\exp(i\gamma\cdot x)$, $x \in (-\pi,\pi)^3$, $\gamma \in \mathbb{Z}^3$. We first bound the sum (3.3) over those multi-indices γ whose norm is larger than a given constant $\rho \geq 2$. It is immediately seen that

$$\begin{aligned}\sum_{|\gamma|>\rho} |\widehat{(n-\tilde{n})}(\gamma)|^2 &\leq \frac{1}{(1+\rho^2)^s} \sum_{|\gamma|>\rho} (1+\gamma\cdot\gamma)^s |\widehat{(n-\tilde{n})}(\gamma)|^2 \\ &\leq \frac{c^2}{\rho^{2s}} \end{aligned} \tag{3.4}$$

because the second factor can be bounded with the help of $\|1-n\|_{H^s} \leq C_n$, $\|1-\tilde{n}\|_{H^s} \leq C_n$. For the estimate of $\|n-\tilde{n}\|_\infty$ we use the Cauchy–Schwarz inequality and obtain

$$\begin{aligned}&\sum_{|\gamma|>\rho} |\widehat{(n-\tilde{n})}(\gamma)| \\ &\leq \left(\sum_{|\gamma|>\rho} (1+\gamma\cdot\gamma)^s |\widehat{(n-\tilde{n})}(\gamma)|^2\right)^{1/2} \left(\sum_{|\gamma|>\rho} \frac{1}{(1+\gamma\cdot\gamma)^s}\right)^{1/2} \\ &\leq \frac{c}{\rho^{s-3/2}}. \end{aligned} \tag{3.5}$$

Here for the second factor we have employed the inequality

$$\sum_{|\gamma|>\rho} \frac{1}{(1+\gamma\cdot\gamma)^s} \leq \frac{c}{\rho^{2s-3}}.$$

Note that c may denote different constants in what follows.

Next, we study the remaining sum of (3.3) over those multi-indices γ with $|\gamma| \leq \rho$. To this end we need the following lemma, which, together with the geometrical optics solutions, is the main tool used to establish a relation between the Fourier coefficients and the operator $S_n - S_{\tilde{n}}$.

LEMMA 3.2. *Assume* $1 < R < R'$. *Moreover,* n, $\tilde{n}$ *are refractive indices with* $\operatorname{supp}(1-n)$, $\operatorname{supp}(1-\tilde{n}) \subset B_1$. *Then, there exists a positive constant* M_3 *(depending only on* k, R, *and* R'*) such that for all solutions* $u \in C^2(B_{R'}) \cap L^2(B_{R'})$ *to* $\Delta u + k^2 n u = 0$ *and all solutions* $\tilde{u} \in C^2(B_{R'}) \cap L^2(B_{R'})$ *to* $\Delta \tilde{u} + k^2 \tilde{n} \tilde{u} = 0$ *in* $B_{R'}$ *the estimate*

$$\left| \int_{B_1} (n - \tilde{n})\, u\, \tilde{u}\, dx \right| \leq M_3 \, \|S_n - S_{\tilde{n}}\|_{L^2(\partial B_R)} \, \|u\|_{L^2(B_{R'})} \, \|\tilde{u}\|_{L^2(B_{R'})} \tag{3.6}$$

holds.

Proof. Given u as in the lemma, we first define a function v by extending $u|_{\overline{B_R}}$ continuously as a radiating solution to the Helmholtz equation in the exterior of B_R, i.e., we set $v|_{\overline{B_R}} = u|_{\overline{B_R}}$ in B_R and $v|_{\mathbb{R}^3 \setminus B_R}$ to be the radiating solution to the Helmholtz equation $\Delta v + k^2 v = 0$ in $\mathbb{R}^3 \setminus \overline{B_R}$ with Dirichlet data $v|_{\partial B_R} = u|_{\partial B_R}$. From Lemma 2.2 we obtain

$$v = S_n \left(\frac{\partial v_-}{\partial \nu} - \frac{\partial v_+}{\partial \nu} \right) \quad \text{on } \partial B_R. \tag{3.7}$$

Proceeding analogously with $\tilde{u}$ to define a function $\tilde{v}$ and establishing the analogue of relation (3.7) for $\tilde{v}$, we use Green's second integral theorem to compute

$$\begin{aligned} &\int_{\partial B_R} \left(\frac{\partial v_-}{\partial \nu} - \frac{\partial v_+}{\partial \nu} \right) (S_n - S_{\tilde{n}}) \left(\frac{\partial \tilde{v}_-}{\partial \nu} - \frac{\partial \tilde{v}_+}{\partial \nu} \right) \, \mathrm{d}s \\ &= \int_{\partial B_R} v \left(\frac{\partial \tilde{v}_-}{\partial \nu} - \frac{\partial \tilde{v}_+}{\partial \nu} \right) \, \mathrm{d}s - \int_{\partial B_R} \tilde{v} \left(\frac{\partial v_-}{\partial \nu} - \frac{\partial v_+}{\partial \nu} \right) \, \mathrm{d}s \\ &= k^2 \int_{B_1} (n - \tilde{n})\, u\, \tilde{u} \, \mathrm{d}x, \end{aligned}$$

and, therefore,

$$\begin{aligned} &\left| \int_{B_1} (n - \tilde{n})\, u\, \tilde{u} \; \mathrm{d}x \right| \\ &\leq \frac{1}{k^2} \, \|S_n - S_{\tilde{n}}\|_{L^2(\partial B_R)} \left\| \frac{\partial v_-}{\partial \nu} - \frac{\partial v_+}{\partial \nu} \right\|_{L^2(\partial B_R)} \left\| \frac{\partial \tilde{v}_-}{\partial \nu} - \frac{\partial \tilde{v}_+}{\partial \nu} \right\|_{L^2(\partial B_R)} . \end{aligned} \tag{3.8}$$

In order to bound the norms of the normal derivatives in (3.8), we apply standard regularity results (Weyl's lemma) for the Helmholtz equation in the shells

$$\Gamma_1 := \left\{ x \in \mathbb{R}^3 \colon \frac{1+R}{2} < |x| < \frac{R'+R}{2} \right\} \subset \Gamma_2 := \{ x \in \mathbb{R}^3 \colon 1 < |x| < R' \}$$

to obtain $\|u\|_{C^2(\Gamma_1)} \leq c \|u\|_{L^2(\Gamma_2)} \leq c \|u\|_{L^2(B_{R'})}$. Hence $\|\partial v_- / \partial \nu\|_{L^2(\partial B_R)}$ can be bounded by $\|u\|_{L^2(B_{R'})}$. Regularity properties for the solution to the exterior boundary value problem for the Helmholtz equation with smooth boundary data $u|_{\partial B_R}$ imply that $\|\partial v_+ / \partial \nu\|_{L^2(\partial B_R)}$ can also be bounded by $\|u\|_{L^2(B_{R'})}$. The same reasoning applies to $\tilde{v}$. Now the lemma follows from inequality (3.8). □

We are now in a position to estimate the Fourier coefficients for the multi-indices $|\gamma| \leq \rho$.

LEMMA 3.3. *Assume $\rho \geq 2$ and C_n are positive constants and n, $\tilde{n}$ are refractive indices satisfying $\|1 - n\|_{H^s}, \|1 - \tilde{n}\|_{H^s} \leq C_n$ and* supp$(1 - n)$, supp$(1 - \tilde{n}) \subset B_1$. *Moreover, suppose $1 < R$ and define $t_0 := \sqrt{M_1^2 + k^2}$ (M_1 being the constant from Lemma* 3.1 *for $R' = 2R$). Then there exists a positive constant M_4 (depending only on s, k, R, and C_n) such that for all $\gamma \in \mathbb{Z}^3$ with $|\gamma| \leq \rho$ and all $t > t_0$ the estimate*

$$\left|\widehat{(n - \tilde{n})}(\gamma)\right| \leq M_4 \left(e^{4R(t+\rho)} \|S_n - S_{\tilde{n}}\|_{L^2(\partial B_R)} + \frac{1}{t} \right) \tag{3.9}$$

holds true.

Proof. For $t > t_0 > k$ and a fixed multi-index $\gamma \in \mathbb{Z}^3$ we choose unit vectors d_1, $d_2 \in \mathbb{R}^3$ such that $d_1 \cdot d_2 = d_1 \cdot \gamma = d_2 \cdot \gamma = 0$ and define the complex vectors

$$\zeta_t := -\frac{1}{2}\, \gamma + i \sqrt{t^2 - k^2 + \frac{|\gamma|^2}{4}}\; d_1 + t\, d_2 \in \mathbb{C}^3, \tag{3.10}$$

$$\tilde{\zeta}_t := -\frac{1}{2}\, \gamma - i \sqrt{t^2 - k^2 + \frac{|\gamma|^2}{4}}\; d_1 - t\, d_2 \ \in \mathbb{C}^3. \tag{3.11}$$

Then the relations $\zeta_t + \tilde{\zeta}_t = -\gamma$, $\zeta_t \cdot \zeta_t = \tilde{\zeta}_t \cdot \tilde{\zeta}_t = k^2$ are satisfied. Furthermore, $t > t_0 = \sqrt{M_1^2 + k^2}$ implies $|\operatorname{Im}(\zeta_t)|, |\operatorname{Im}(\tilde{\zeta}_t)| \geq M_1$. Therefore, according to Lemma 3.1, there exist the geometrical optics solutions

$$U(x, \zeta_t) = e^{i\zeta_t \cdot x}(1 + v(x, \zeta_t)), \quad x \in B_{2R},$$

to $\Delta u + k^2 n u = 0$ and

$$\tilde{U}(x, \tilde{\zeta}_t) = e^{i\tilde{\zeta}_t \cdot x}(1 + \tilde{v}(x, \tilde{\zeta}_t)), \quad x \in B_{2R},$$

to $\Delta \tilde{u} + k^2 \tilde{n} \tilde{u} = 0$. Using Lemma 3.1 once more, we infer

$$U(x, \zeta_t)\, \tilde{U}(x, \tilde{\zeta}_t) = e^{-i\gamma \cdot x}(1 + p(x, t)),$$

where the L^1 norm of the remainder

$$p(x, t) = v(x, \zeta_t) + \tilde{v}(x, \tilde{\zeta}_t) + v(x, \zeta_t)\, \tilde{v}(x, \tilde{\zeta}_t)$$

is bounded by

$$\int_{B_1} |p(x, t)| \ \mathrm{d}x \leq \frac{c}{t} \tag{3.12}$$

with a suitable constant c. Then we compute

$$\begin{aligned}
&\left|\widehat{(n - \tilde{n})}(\gamma)\right| \\
&= \frac{1}{(2\pi)^{3/2}} \left| \int_{B_1} (n - \tilde{n})(x)\, e^{-i\gamma \cdot x} \ \mathrm{d}x \right| \\
&= \frac{1}{(2\pi)^{3/2}} \left| \int_{B_1} (n - \tilde{n})(x)\, U(x, \zeta_t)\, \tilde{U}(x, \tilde{\zeta}_t) \ \mathrm{d}x \right. \\
&\qquad\qquad \left. - \int_{B_1} (n - \tilde{n})(x)\, e^{-i\gamma \cdot x}\, p(x, t) \ \mathrm{d}x \right| \\
&\leq c' \, \|S_n - S_{\tilde{n}}\|_{L^2(\partial B_R)}\, \|U(\cdot, \zeta_t)\|_{L^2(B_{2R})}\, \|\tilde{U}(\cdot, \tilde{\zeta}_t)\|_{L^2(B_{2R})} + \frac{c''}{t} \\
&\leq M_4 \left(e^{4R(t+\rho)} \|S_n - S_{\tilde{n}}\|_{L^2(\partial B_R)} + \frac{1}{t} \right),
\end{aligned} \tag{3.13}$$

where we have used the estimates (3.6) and (3.12) for the first inequality. For the last line in (3.13) we note that

$$\begin{aligned}\|U(\cdot,\zeta_t)\|_{L^2(B_{2R})} &= \|e^{i\zeta_t\cdot x}\,(1+v(\cdot,\zeta_t))\|_{L^2(B_{2R})}\\ &\le \|e^{i\zeta_t\cdot x}\|_{\infty,B_{2R}}\|1+v(\cdot,\zeta_t)\|_{L^2(B_{2R})}\\ &\le c'''e^{2R(t+\rho)}\end{aligned}$$

for all $t \ge t_0$ and all $\gamma \in \mathbb{Z}^3$ with $|\gamma| \le \rho$ because of $|\operatorname{Im}(\zeta_t)| \le t + |\gamma| \le t+\rho$. Together with the analogous estimate for $\|\tilde{U}(\cdot,\tilde{\zeta}_t)\|_{L^2(B_{2R})}$, we obtain (3.13), i.e., we have proved inequality (3.9). □

Finally, we can prove Theorem 1.1 by choosing the parameters ρ and t in (3.4), (3.5), and (3.9) appropriately.

Proof of Theorem 1.1. In view of the Fourier expansion (3.3) for $n-\tilde{n}$, Parseval's equation, the inequalities (3.4) and (3.9), and the fact that there are less than $c\rho^3$ multi-indices $\gamma \in \mathbb{Z}^3$ with $|\gamma| \le \rho$, we have

$$\begin{aligned}\|n-\tilde{n}\|^2_{L^2(B_1)} &\le \frac{c^2}{\rho^{2s}} + \left(\sum_{|\gamma|\le\rho} |\widehat{(n-\tilde{n})}(\gamma)|\right)^2\\ &\le \frac{c^2}{\rho^{2s}} + \left(c\rho^3\left[e^{4R(t+\rho)}\|S_n - S_{\tilde{n}}\|_{L^2(\partial B_R)} + \frac{1}{t}\right]\right)^2\\ &\le c\left(e^{(4R+1)(t+\rho)}\|S_n - S_{\tilde{n}}\|_{L^2(\partial B_R)} + \frac{\rho^3}{t} + \frac{1}{\rho^s}\right)^2, \qquad (3.14)\end{aligned}$$

where we have also used $\rho^3 e^{4R(t+\rho)} \le e^{(4R+1)(t+\rho)}$ (due to $\rho^3 \le 6e^\rho$) in the last line. Note that the value of the constants c will change during the proof. For the maximum norm, in view of (3.3) and (3.5), we compute

$$\begin{aligned}\|n-\tilde{n}\|_\infty &\le c\sum_{|\gamma|\le\rho} |\widehat{(n-\tilde{n})}(\gamma)| + \frac{c}{\rho^{s-3/2}}\\ &\le c\rho^3\left(e^{4R(t+\rho)}\|S_n - S_{\tilde{n}}\|_{L^2(\partial B_R)} + \frac{1}{t}\right) + \frac{c}{\rho^{s-3/2}}\\ &\le c\left(e^{(4R+1)(t+\rho)}\|S_n - S_{\tilde{n}}\|_{L^2(\partial B_R)} + \frac{\rho^3}{t} + \frac{1}{\rho^{s-3/2}}\right). \qquad (3.15)\end{aligned}$$

For $t > t_0 + 2^{s+3}$ (t_0 from Lemma 3.3) and $\rho := t^{1/(s+3)}$ the condition $\rho \ge 2$ is satisfied and inequality (3.14) becomes

$$\|n-\tilde{n}\|_{L^2(B_1)} \le c\left(e^{(8R+2)t}\|w^{\mathrm{s}}_n - w^{\mathrm{s}}_{\tilde{n}}\|_{L^2(\partial B_R\times\partial B_R)} + \frac{2}{t^{s/(s+3)}}\right) \qquad (3.16)$$

because of $\rho = t^{1/(s+3)} \le t$ and $\|S_n - S_{\tilde{n}}\|_{L^2(\partial B_R)} \le \|w^{\mathrm{s}}_n - w^{\mathrm{s}}_{\tilde{n}}\|_{L^2(\partial B_R\times\partial B_R)}$.

We may restrict our attention to the case that $\|w^{\mathrm{s}}_n - w^{\mathrm{s}}_{\tilde{n}}\|_{L^2}$ is sufficiently small such that

$$t := -\frac{1}{s+3}\;\frac{3}{8R+2}\;\ln\left(\|w^{\mathrm{s}}_n - w^{\mathrm{s}}_{\tilde{n}}\|_{L^2}\right) > t_0 + 2^{s+3}.$$

Otherwise, the bound on $\|n-\tilde{n}\|_\infty$ from the a priori information about n and $\tilde{n}$ and the monotonicity of $[-\ln^-(t)]^{-p}$, $0<p<1$, imply

$$\|n-\tilde{n}\|_{L^2(B_1)} \le c\,\frac{[-\ln^-(\|w^{\mathrm{s}}_n - w^{\mathrm{s}}_{\tilde{n}}\|_{L^2})]^{-p}}{[(t_0+2^{s+3})\,\frac{s+3}{3}\,(8R+2)]^{-p}}.$$

Inserting t into (3.16) yields the inequality

$$\begin{aligned}\|n-\tilde{n}\|_{L^2(B_1)} &\leq c\left\{\|w_n^{\mathrm{s}}-w_{\tilde{n}}^{\mathrm{s}}\|_{L^2}^{s/(s+3)}+(-\ln\|w_n^{\mathrm{s}}-w_{\tilde{n}}^{\mathrm{s}}\|_{L^2})^{-s/(s+3)}\right\}\\ &\leq c\left[-\ln\left(\|w_n^{\mathrm{s}}-w_{\tilde{n}}^{\mathrm{s}}\|_{L^2}\right)\right]^{-s/(s+3)}\end{aligned}$$

because $x\leq(-\ln(x))^{-1}$ for $0<x<1$.

The analogous analysis with $\rho:=t^{2/(2s+3)}$ and

$$t:=-\frac{1}{2s+3}\,\frac{3}{4R+1}\,\ln\left(\|w_n^{\mathrm{s}}-w_{\tilde{n}}^{\mathrm{s}}\|_{L^2}\right)>t_0+2^{s+3}$$

reveals

$$\begin{aligned}\|n-\tilde{n}\|_\infty &\leq c\left\{\|w_n^{\mathrm{s}}-w_{\tilde{n}}^{\mathrm{s}}\|_{L^2}^{(2s-3)/(2s+3)}+(-\ln\|w_n^{\mathrm{s}}-w_{\tilde{n}}^{\mathrm{s}}\|_{L^2})^{-(2s-3)/(2s+3)}\right\}\\ &\leq c\left[-\ln\left(\|w_n^{\mathrm{s}}-w_{\tilde{n}}^{\mathrm{s}}\|_{L^2}\right)\right]^{-(2s-3)/(2s+3)}.\end{aligned}$$

This finishes the proof. □

4. Estimating the refractive index with the help of far field data (proof of Theorem 1.2). Our proof of Theorem 1.2 is based on the following abstract linear stability result. It has been proved in [11, 14] for bounded operators. For the convenience of the reader we include the short proof for compact operators which is sufficient for our purposes.

LEMMA 4.1. *Let X,Y be Hilbert spaces, and let $T:X\to Y$ be a compact linear operator. Moreover, assume that $f\in C([0,\|T^*T\|])$ is monotonically increasing with $f(0)=0$ and that the function*

$$\phi_f:[0,f(\|T^*T\|)^2]\to[0,\|T^*T\|f(\|T^*T\|)^2]$$

defined by $\phi_f(\xi):=\xi\cdot(f\cdot f)^{-1}(\xi)$ is convex. Then the source condition

$$w=f(T^*T)v,\qquad\|v\|\leq\rho,\tag{4.1}$$

implies the stability estimate

$$\|w\|^2\leq\rho^2\phi_f^{-1}\left(\frac{\|Tw\|^2}{\rho^2}\right).\tag{4.2}$$

Proof. We can assume that x_μ, $\mu\in\mathcal{M}$, is a Hilbert basis of X with $T^*Tx_\mu=\lambda_\mu x_\mu$, and $\lambda_\mu\geq0$ for all $\mu\in\mathcal{M}$. Replacing w by $(1/\rho)w$, we see that we can also assume that $\rho=1$. If $v=\sum_{\mu\in\mathcal{M}}\alpha_\mu x_\mu$, then $w=\sum_{\mu\in\mathcal{M}}f(\lambda_\mu)\alpha_\mu x_\mu$. We first assume that $\|v\|^2=\sum_{\mu\in\mathcal{M}}|\alpha_\mu|^2=1$. Then Jensen's inequality yields

$$\begin{aligned}\phi_f(\|w\|^2)=\phi_f\left(\sum_{\mu\in\mathcal{M}}(f(\lambda_\mu))^2|\alpha_\mu|^2\right)&\leq\sum_{\mu\in\mathcal{M}}\phi_f((f\cdot f)(\lambda_\mu))\,|\alpha_\mu|^2\\ =\sum_{\mu\in\mathcal{M}}(f\cdot f)(\lambda_\mu)\,\lambda_\mu\,|\alpha_\mu|^2&=\|Tw\|^2.\end{aligned}$$

Since ϕ_f is convex and $\phi_f(0)=0$, this inequality is also valid for $\|v\|<1$. Applying ϕ_f^{-1} to both sides of the inequality yields the assertion since ϕ_f^{-1} is increasing under the given assumptions. □

We want to apply this result to the operator T which maps the difference $w_n - w_{\tilde{n}} = w_n^{\mathrm{s}} - w_{\tilde{n}}^{\mathrm{s}}$ of two Green functions on $\partial B_{2R} \times \partial B_{2R}$ to the corresponding difference of far field patterns $u_n^\infty - u_{\tilde{n}}^\infty$. For the function f we choose $f = g_\theta$, $0 < \theta < 1$, defined by

$$g_\theta(\lambda) := \begin{cases} \exp(-\frac{1}{2}(-\ln\lambda)^\theta), & 0 < \lambda \le \exp(-1), \\ 0, & \lambda = 0. \end{cases} \tag{4.3}$$

The corresponding functions $\phi_\theta := \phi_{g_\theta}$ and their second derivatives are given by

$$\begin{aligned} \phi_\theta(\xi) &= \xi \exp(-(-\ln\xi)^{\frac{1}{\theta}}), \qquad 0 < \xi \le \exp(-1), \\ \phi_\theta''(\xi) &= \exp\left(-(-\ln\xi)^{\frac{1}{\theta}}\right)\left(\frac{(-\ln\xi)^{\frac{1}{\theta}-2}}{\theta\xi}\right)\left(1 - \ln\xi + \frac{(-\ln\xi)^{\frac{1}{\theta}}-1}{\theta}\right). \end{aligned}$$

It is obvious that the functions g_θ are continuous and monotonically increasing and that $\phi_\theta''(\xi) > 0$ for $0 < \xi < \exp(-1)$ and $0 < \theta < 1$.

With the help of the orthonormal basis

$$Y_{r,\mu}(x,y) = \frac{1}{r^2}\, Y_{l_1}^{m_1}\left(\frac{x}{|x|}\right) Y_{l_2}^{m_2}\left(\frac{y}{|y|}\right), \quad \mu = (l_1, m_1, l_2, m_2) \in \mathcal{M},$$

of $L^2(\partial B_r \times \partial B_r)$ and the basis $Y_\mu := Y_{1,\mu}$ in $L^2(S^2 \times S^2)$, we infer from the series expansions in Lemma 2.3 that T defined by

$$Y_{2R,\mu} \mapsto -\frac{\pi}{k^2R^2}\, i^{l_2-l_1}\, \frac{1}{h_{l_1}^{(1)}(2kR) h_{l_2}^{(1)}(2kR)}\, Y_\mu \tag{4.4}$$

maps $(w_n - w_{\tilde{n}})|_{\partial B_{2R} \times \partial B_{2R}}$ to $u_n^\infty - u_{\tilde{n}}^\infty$. Because g_θ is defined only on $[0, \exp(-1)]$, we introduce the scaling factor $\omega := \|T\| \exp(1/2)$ and use the operator $\omega^{-1}T$. In order to apply Lemma 4.1, we must check that a function $w_n - w_{\tilde{n}}$ from which it is known a priori that it originates from refractive indices n, $\tilde{n}$ such that $\|1-n\|_{H^s}, \|1-\tilde{n}\|_{H^s} \le C_n$, $\operatorname{supp}(1-n)$, $\operatorname{supp}(1-\tilde{n}) \subset B_1$, satisfies the source condition, i.e., there is a constant $\rho > 0$ such that $(w_n - w_{\tilde{n}})|_{\partial B_{2R} \times \partial B_{2R}} = g_\theta(\omega^{-2}T^*T)v$ for an appropriate $v \in L^2(\partial B_{2R} \times \partial B_{2R})$ with $\|v\| \le \rho$. We do this in the following lemma.

Lemma 4.2. *Let $1 < R$, $C_n > 0$, and $0 < \theta < 1$ be given. Furthermore, define $\omega := \|T\| \exp(1/2)$, where T denotes the operator from* (4.4). *Then there exists a constant $\rho > 0$ such that all $(w_n - w_{\tilde{n}})|_{\partial B_{2R} \times \partial B_{2R}}$ originating from refractive indices n, $\tilde{n}$ such that $\|1-n\|_{H^s}, \|1-\tilde{n}\|_{H^s} \le C_n$, $\operatorname{supp}(1-n)$, $\operatorname{supp}(1-\tilde{n}) \subset B_1$, satisfy the source condition $(w_n - w_{\tilde{n}})|_{\partial B_{2R} \times \partial B_{2R}} = g_\theta(\omega^{-2}T^*T)v$ for an appropriate $v \in L^2(\partial B_{2R} \times \partial B_{2R})$ with $\|v\| \le \rho$.*

Proof. We know from Lemma 2.3 that

$$(w_n - w_{\tilde{n}})(r\hat{x}, r\hat{y}) = \sum_{\mu \in \mathcal{M}} \alpha_\mu\, r^2\, h_{l_1}^{(1)}(kr)\, h_{l_2}^{(1)}(kr)\, Y_{r,\mu}(r\hat{x}, r\hat{y})$$

for all $r \ge R$ with suitable coefficients α_μ. For convenience we set

$$\gamma_\mu := -\pi(kR\, i^{l_1} h_{l_1}^{(1)}(2kR))^{-1}(kR(-i)^{l_2}\, h_{l_2}^{(1)}(2kR))^{-1}$$

and

$$\delta_\mu = 4\, \frac{h_{l_1}^{(1)}(2kR)}{h_{l_1}^{(1)}(kR)}\, \frac{h_{l_2}^{(1)}(2kR)}{h_{l_2}^{(1)}(kR)}.$$

Note that $|\gamma_\mu|$ are the singular values of T. If we can verify that the function

$$v := \sum_{\mu\in\mathcal{M}} \alpha_\mu \; R^2 \, h^{(1)}_{l_1}(kR) \, h^{(1)}_{l_2}(kR) \; \frac{\delta_\mu}{g_\theta(|\gamma_\mu|^2/\omega^2)} \; Y_{2R,\mu}$$

belongs to $L^2(\partial B_{2R} \times \partial B_{2R})$ and $\|v\| \leq \rho$, a straightforward computation shows $g_\theta(\omega^{-2}T^*T)v = w_n - w_{\tilde{n}}$ on $\partial B_{2R} \times \partial B_{2R}$, and we have proved the source condition.

Since we know from Lemma 2.1(b) that

$$\sum_{\mu\in\mathcal{M}} \left|\alpha_\mu \; R^2 \, h^{(1)}_{l_1}(kR) \, h^{(1)}_{l_2}(kR)\right|^2 = \|(w_n - w_{\tilde{n}})|_{\partial B_R \times \partial B_R}\|^2 \leq c^2$$

for all refractive indices satisfying the assumptions, it suffices to prove

$$\sup_{\mu\in\mathcal{M}} \frac{|\delta_\mu|}{g_\theta(|\gamma_\mu/\omega|^2)} \leq c, \tag{4.5}$$

or in other words that the supremum in (4.5) is finite. (The variable c denotes various constants during the proof.)

To see this, we use the asymptotic formulae

$$\begin{aligned} \frac{1}{c}\left(\frac{l}{kRe}\right)^l \leq |h^{(1)}_l(2kR)| \leq c\left(\frac{l}{kRe}\right)^l, \qquad l \to \infty, \\ \frac{h^{(1)}_l(2kR)}{h^{(1)}_l(kR)} = \frac{1}{2^{l+1}}\;\left(1 + \mathrm{O}\left(\frac{1}{l}\right)\right), \qquad l \to \infty, \end{aligned} \tag{4.6}$$

which follow from the series representation of the spherical Hankel functions (cf. [3]). Therefore, we have

$$|\delta_\mu| \leq c\, 2^{-l_1-l_2}. \tag{4.7}$$

Using the monotonicity of g_θ from (4.6) and the inequality $(a+b)^\theta \leq a^\theta + b^\theta$, which holds for $a, b \geq 0$ and $0 \leq \theta \leq 1$ (for a proof consider the case $a + b = 1$), we obtain the estimate

$$\begin{aligned} g_\theta(|\gamma_\mu/\omega|^2)^{-2} &\leq g_\theta\left(c^2\left(\frac{ekR}{l_1}\right)^{2l_1}\left(\frac{ekR}{l_2}\right)^{2l_2}\right)^{-2} \\ &= \exp\left(\left\{\ln\left(c^2\left(\frac{l_1}{ekR}\right)^{2l_1}\left(\frac{l_2}{ekR}\right)^{2l_2}\right)\right\}^\theta\right) \\ &\leq \exp\left(\left\{\ln c + 2l_1 \ln\frac{l_1}{ekR}\right\}^\theta + \left\{\ln c + 2l_2 \ln\frac{l_2}{ekR}\right\}^\theta\right). \end{aligned}$$

Since $\lim_{l\to\infty}(\{\ln c + 2l\ln(l/(ekR))\}^\theta - 2l\ln 2) = -\infty$, it follows from (4.7) that

$$\frac{|\delta_\mu|^2}{g_\theta(|\gamma_\mu/\omega|^2)^2} \leq c$$

for all $\mu \in \mathcal{M}$. This shows (4.5) and finishes the proof. $\square$

Proof of Theorem 1.2. Lemma 4.2 yields

$$(w_n - w_{\tilde n})|_{\partial B_{2R}\times\partial B_{2R}} = g_\theta(\omega^{-2}T^*T)v, \quad \|v\| \le \rho.$$

Hence, using $T((w_n - w_{\tilde n})|_{\partial B_{2R}\times\partial B_{2R}}) = u_n^\infty - u_{\tilde n}^\infty$, according to Lemma 4.1 we have

$$\begin{aligned}\|(w_n - w_{\tilde n})|_{\partial B_{2R}\times\partial B_{2R}}\|^2 &\le \rho^2\phi_\theta^{-1}\left(\frac{\|T((w_n - w_{\tilde n})|_{\partial B_{2R}\times\partial B_{2R}})\|^2}{\omega^2\rho^2}\right)\\ &= \rho^2\phi_\theta^{-1}\left(\frac{\|u_n^\infty - u_{\tilde n}^\infty\|^2}{\omega^2\rho^2}\right).\end{aligned}$$

To estimate ϕ_θ^{-1}, we write $\phi_\theta(\xi) = \exp(-\psi(-\ln\xi))$ with $\psi(x) := x + x^{\frac{1}{\theta}}$. Obviously, $\psi(x) \le 2x^{\frac{1}{\theta}}$ for $x \ge 1$, so $\psi^{-1}(y) \ge \left(\frac{y}{2}\right)^\theta$ for $y \ge 2$. It follows that

$$\phi_\theta^{-1}(\lambda) = \exp(-\psi^{-1}(-\ln\lambda)) \le \exp\left(-\left(\frac{-\ln\lambda}{2}\right)^\theta\right)$$

for $0 < \lambda \le \exp(-2)$, and we get

$$\|w_n^{\mathrm{s}} - w_{\tilde n}^{\mathrm{s}}\|^2_{L^2(\partial B_{2R}\times\partial B_{2R})} \le \rho^2\exp\left(-\left(-\ln\frac{\|u_n^\infty - u_{\tilde n}^\infty\|_{L^2(S^2\times S^2)}}{\omega\rho}\right)^\theta\right)$$

for sufficiently small $\|u_n^\infty - u_{\tilde n}^\infty\|_{L^2(S^2\times S^2)}$. Inserting this into (1.3) and choosing θ such that $\theta\frac{s}{s+3} = \frac{s}{s+3} - \epsilon$ yield the assertion. By adjusting the constant, similarly to the proof of Theorem 1.1, this inequality holds for all refractive indices satisfying the assumptions. The estimate for $\|n - \tilde n\|_\infty$ is derived analogously. ∎

5. Applications. In this final section we discuss the numerical solution of the inverse problem to recover the refractive index $n(x) = 1 - a(x)$ from measurements of the far field pattern $u^\infty(\hat x, d)$ by iterative regularization methods. Introducing the Hilbert spaces $X = H_0^s(B_1)$, $Y = L^2(S^2\times S^2)$ and the operator $F : X \to Y$, $F(a) := u_{1-a}^\infty$, this problem can be formulated as a nonlinear operator equation:

$$F(a) = u^\infty. \tag{5.1}$$

Usually the measured data will be contaminated by noise. We use the notation $u^{\infty,\delta}$ with $\|u^{\infty,\delta} - u^\infty\| \le \delta$ for noisy data with noise level δ.

A number of methods have been developed to solve general problems of the form (5.1). Among the most popular are the *Landweber iteration*

$$a_{j+1}^\delta := a_j^\delta + \mu F'[a_j^\delta]^*(u^{\infty,\delta} - F(a_j^\delta)) \tag{5.2}$$

($\mu \le \|F'[a]\|^{-1}$ for a in a neighborhood of the exact solution $a^\dagger$), the *Levenberg–Marquardt algorithm*

$$a_{j+1}^\delta := a_j^\delta + (\alpha_j I + F'[a_j^\delta]^*F'[a_j^\delta])^{-1}F'[a_j^\delta]^*(u^{\infty,\delta} - F(a_j^\delta)), \tag{5.3}$$

and the *iteratively regularized Gauß–Newton method* (IRGNM)

(5.4)
$$a_{j+1}^\delta := a_0 + (\alpha_j I + F'[a_j^\delta]^*F'[a_j^\delta])^{-1}F'[a_j^\delta]^*\left(u^{\infty,\delta} - F(a_j^\delta) + F'[a_j^\delta](a_j^\delta - a_0)\right).$$

For the last two methods there are different strategies for choosing the regularization parameters α_j, e.g., $\alpha_j = 2^{-j}\alpha_0$.

An important issue is the choice of the stopping index. The most well-known stopping rule is *Morozov's discrepancy principle*, which consists in stopping the iteration at the first index $J = J(\delta, u^{\infty,\delta})$ such that

$$\|F(a_J^\delta) - u^{\infty,\delta}\| \leq \tau\delta \tag{5.5}$$

with some fixed constant $\tau \geq 1$.

We call an iterative method $a_{j+1}^\delta := \Psi_j(a_j, \ldots, a_0, u^{\infty,\delta})$ together with a stopping rule $J(\delta, u^{\infty,\delta})$ an *iterative regularization method* for F if for all initial guesses a_0 in a neighborhood of $a^\dagger$ the iterates a_{j+1}^δ, $j \leq J(\delta, u^{\infty,\delta})$ are well defined and

$$\lim_{j\to\infty} a_j^0 = a^\dagger, \tag{5.6}$$

$$\lim_{\delta\to 0} \sup_{\|u^{\infty,\delta}-u^\infty\|\leq\delta} \|a_{J(\delta,u^{\infty,\delta})}^\delta - a^\dagger\| = 0. \tag{5.7}$$

It has been shown that under certain conditions restricting the degree of nonlinearity of the operator F, the Landweber iteration (Hanke, Neubauer, and Scherzer [9]), the Levenberg–Marquardt algorithm (Hanke [8]), and the IRGNM (Blaschke/Kaltenbacher, Neubauer, Scherzer [2]), together with the discrepancy principle (5.5), are regularization methods in the sense of the definition above. It must be mentioned that none of these nonlinearity conditions has been verified for the inverse acoustic medium scattering problem yet.

Since it is well known that the convergence in (5.7) can be *arbitrarily slow* for any ill-posed problem (cf. [5, Proposition 3.11]), one is interested in proving estimates on the convergence rate under additional a priori assumptions. In the literature, assumptions of the form

$$a^\dagger - a_0 = f(F'[a^\dagger]^* F'[a^\dagger])v, \qquad \|v\| \leq \rho, \tag{5.8}$$

have been studied intensively. These conditions, called *nonlinear source conditions*, are nonlinear analogues to (4.1). It turns out that for many exponentially ill-posed problems source conditions (5.8) with $f = f_p$ $(p > 0)$

$$f_p(\lambda) := \begin{cases} (-\ln\lambda)^{-p}, & 0 < \lambda \leq \exp(-1), \\ 0, & \lambda = 0 \end{cases} \tag{5.9}$$

(so-called *logarithmic source conditions*), have natural interpretations as smoothness conditions in Sobolev spaces (cf. [11, 13]). The usual Hölder-type source conditions (5.8) with $f(\lambda) = \lambda^\mu$, $\mu > 0$, are appropriate for mildly ill-posed problems but typically far too restrictive for exponentially ill-posed problems. Under the logarithmic source condition (5.8), (5.9), convergence rates of the form

$$\|a_{J(\delta,u^{\infty,\delta})}^\delta - a^\dagger\| = \mathrm{O}\left(f_p(\delta)\right)$$

have been established for the IRGNM [10] and for the Landweber iteration [4].

Unfortunately, it seems that logarithmic source conditions for the inverse acoustic medium scattering problem do not have an equivalent formulation with the help of known Sobolev spaces. However, Theorem 1.2 gives us a sufficient condition for convergence rates under the simple assumption that the exact solution $n^\dagger = 1 - a^\dagger$ and the initial guess $n_0 := 1 - a_0$ belong to $H_0^s(B_1)$.

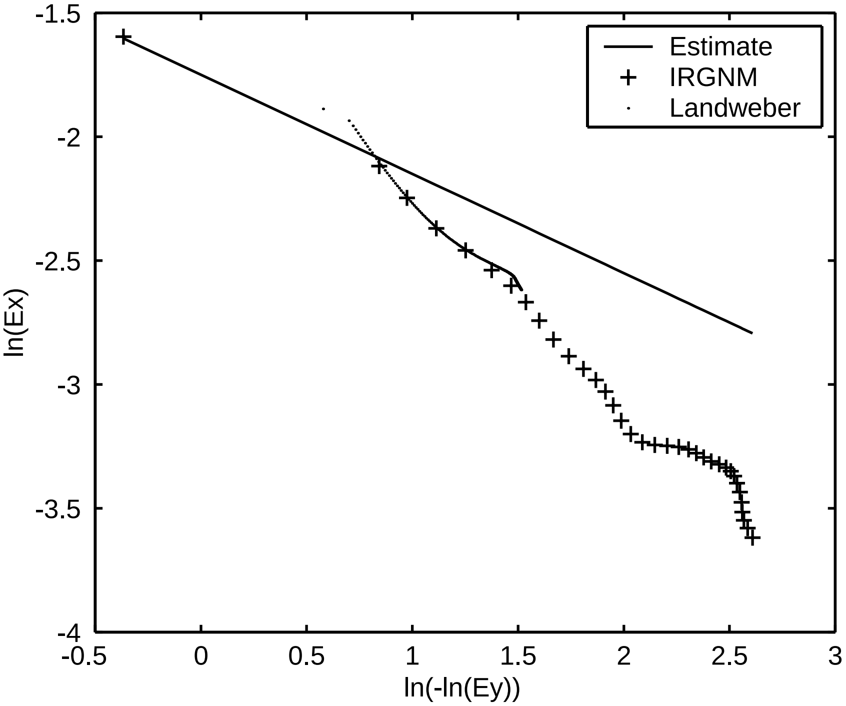


FIG. 5.1. *Plot of* $Ex := \|a_j^0 - a^\dagger\|_{L^2}$ *on a logarithmic scale over* $Ey := \|u^\infty_{1-a_j^0} - u^\infty_{1-a^\dagger}\|_{L^2}$ *on a double logarithmic scale.*

COROLLARY 5.1. *Assume that* $a^\dagger, a_0 \in H^s(B_1)$. *Let* a_j^δ *denote the iterates generated by an iterative regularization method for* F *which converges for the initial guess* a_0, *and assume that the stopping index* J *is chosen according to* (5.5). *Then for any* $\epsilon > 0$

$$\|a_j^0 - a^\dagger\|_{L^2(B_1)} = \mathrm{O}\left(-\ln\|u^\infty_{1-a_j^0} - u^\infty_{1-a^\dagger}\|\right)^{-\frac{s}{s+3}+\epsilon}, \qquad j \to \infty, \tag{5.10}$$

$$\|a^\delta_{J(\delta,u^{\infty,\delta})} - a^\dagger\|_{L^2(B_1)} = \mathrm{O}\left((-\ln\delta)^{-\frac{s}{s+3}+\epsilon}\right), \qquad \delta \to 0. \tag{5.11}$$

Proof. Due to (5.6) we know that there is a constant C_n such that $\|a_j^0\|_{H^s} \le C_n$ for all j. Hence Theorem 1.2 with $n = 1 - a^\dagger$ and $\tilde{n} = 1 - a_j^0$ yields (5.10). Analogously, (5.11) follows from (5.7), (5.5), and Theorem 1.2. □

In Figure 5.1 the actual values of $\|a_j^0 - a^\dagger\|_{L^2}$ and $\|u^\infty_{1-a_j^0} - u^\infty_{1-a^\dagger}\|_{L^2}$ for the IRGNM and the Landweber iteration are compared to the estimate (5.10). We used the Hilbert space $X = H_0^2(B_1)$, i.e., $s = 2$. To implement the IRGNM we solved the equation system involving $\alpha_j I + F'[a_j^\delta]^* F'[a_j^\delta]$ by the conjugate gradient method using a special preconditioner (cf. [12]; our test problem is taken from section 3 of this paper). We performed 500 Landweber iterations and 20 Newton iterations using exact synthetic data, which were created using a modified algorithm with a much finer discretization in order to avoid an inverse crime. The constant in (5.10) was chosen to fit the data. The picture shows that our estimate is of the right form. The exponent $-\frac{s}{s+3} = -\frac{2}{5}$ is fairly close to our experimental results, but it seems to be a little too pessimistic. In fact, using (3.9), we cannot get an exponent greater than 1 even for a single Fourier coefficient. As long as estimate (3.9) is used, we cannot expect to obtain a better stability result than

$$\|n-\tilde{n}\|_{L^2(B_1)} \leq C\left[-\ln^-\left(\|w_n^{\mathrm{s}}-w_{\tilde{n}}^{\mathrm{s}}\|_{L^2(\partial B_R\times\partial B_R)}\right)\right]^{-1}.$$

To our knowledge it is still open whether there is an example which shows that no better estimate is possible or whether new ideas could improve (3.9) and then the stability estimate.

REFERENCES

[1] S. Alessandrini, *Stable determination of conductivity by boundary measurements*, Appl. Anal., 27 (1988), pp. 153–172.

[2] B. Blaschke, A. Neubauer, and O. Scherzer, *On convergence rates for the iteratively regularized Gauß–Newton method*, IMA J. Numer. Anal., 17 (1997), pp. 421–436.

[3] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer-Verlag, Berlin, Heidelberg, New York, 1998.

[4] P. Deuflhard, H. Engl, and O. Scherzer, *A convergence analysis of iterative methods for the solution of nonlinear ill-posed problems under affinely invariant conditions*, Inverse Problems, 14 (1998), pp. 1081–1106.

[5] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, Dordrecht, Boston, London, 1996.

[6] P. Hähner, *A periodic Faddeev-type solution operator*, J. Differential Equations, 128 (1996), pp. 300–308.

[7] P. Hähner, *On Acoustic, Electromagnetic, and Elastic Scattering Problems in Inhomogeneous Media*, Habilitation thesis, Universität Göttingen, Göttingen, Germany, 1998.

[8] M. Hanke, *A regularizing Levenberg-Marquardt scheme, with applications to inverse groundwater filtration problems*, Inverse Problems, 13 (1997), pp. 79–95.

[9] M. Hanke, A. Neubauer, and O. Scherzer, *A convergence analysis of the Landweber iteration for nonlinear ill-posed problems*, Numer. Math., 72 (1995), pp. 21–37.

[10] T. Hohage, *Logarithmic convergence rates of the iteratively regularized Gauß-Newton method for an inverse potential and an inverse scattering problem*, Inverse Problems, 13 (1997), pp. 1279–1299.

[11] T. Hohage, *Iterative Methods in Inverse Obstacle Scattering: Regularization Theory of Linear and Nonlinear Exponentially Ill-Posed Problems*, Ph.D. thesis, University of Linz, Linz, Austria, 1999. Universitätsverlag Rudolf Trauner, Linz, Austria, 1999.

[12] T. Hohage, *On the Numerical Solution of a* 3*D Inverse Medium Scattering Problem*, Technical report 00-27, SFB F013, University of Linz, Linz, Austria, 2000.

[13] T. Hohage, *Regularization of exponentially ill-posed problems*, Numer. Funct. Anal. Optim., 21 (2000), pp. 439–464.

[14] B. A. Mair, *Tikhonov regularization for finitely and infinitely smoothing operators*, SIAM J. Math. Anal., 25 (1994), pp. 135–147.

[15] A. Nachman, *Reconstructions from boundary measurements*, Ann. of Math. (2), 128 (1988), pp. 531–576.

[16] R. Pike and P. Sabatier, eds., *Scattering*, Academic Press, London, 2001.

[17] P. Stefanov, *Stability of the inverse problem in potential scattering at fixed energy*, Ann. Inst. Fourier (Grenoble) 40 (1990), pp. 867–884.

[18] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem*, Ann. of Math. (2), 125 (1987), pp. 153–169.

SIAM J. MATH. ANAL.
Vol. 33, No. 3, pp. 686–698

INEQUALITIES OF DUFFIN–SCHAEFFER TYPE[*]

GENO NIKOLOV[†]

Abstract. We prove here that if an algebraic polynomial f of degree at most n has smaller absolute values than T_n (the nth Chebyshev polynomial of the first kind) at arbitrary $n+1$ points in $[-1,1]$, which interlace with the zeros of T_n, then the uniform norm of f' in $[-1,1]$ is smaller than n^2. This is an extension of a classical result obtained by Duffin and Schaeffer.

Key words. Chebyshev polynomial, Markov inequality, numerical differentiation

AMS subject classifications. 41A05, 41A17

PII. S0036141000373914

1. Introduction and statement of the result.

Denote by π_n the class of algebraic polynomials of degree at most n and by $\|\cdot\|$ the supremum norm in $[-1,1]$. The classical inequality of the brothers Markov [8], [9] asserts that, among all $f \in \pi_n$ satisfying

$$\|f\| \le 1, \tag{1}$$

the Chebyshev polynomial of the first kind $T_n(x) = \cos n \arccos x$ has the greatest norm of its kth derivative ($k = 1, \dots, n$). For more than one century the inequality of the Markov brothers has been a challenge for many mathematicians and an object of various generalizations (see, e.g., [3], [4, Chapter 4], [7, Chapter 2]). Inequalities of the Markov type, relating norms of a polynomial and its derivatives, have found numerous applications in both approximation theory and numerical analysis.

One of the most striking extensions of the classical Markov inequality was found by Duffin and Schaeffer [5], who showed that the extremal property of T_n persists under a weaker assumption than (1). Namely, Duffin and Schaeffer proved that T_n still has the largest uniform norm of its kth derivative in the wider class of polynomials from π_n, satisfying

$$|f(\cos(\nu\pi/n))| \le 1,\ \ \nu = 0, \dots, n. \tag{2}$$

(Actually, Duffin and Schaeffer proved a more general result, including an inequality on a strip in the complex plane, but it does not fall in the frame of the present paper.) Let us mention that (2) is a more natural restriction than (1) since to bound a norm of a polynomial of degree n it suffices to impose restrictions on its absolute value at $n+1$ points only.

The points

$$\eta_\nu := \cos(\nu\pi/n),\ \ \nu = 0, \dots, n,$$

are the local extremum points for T_n in $[-1,1]$ and $|T_n(\eta_\nu)| = 1$. Thus the result of Duffin and Schaeffer may be viewed as a comparison-type theorem: the inequality

[*] Received by the editors June 16, 2000; accepted for publication (in revised form) April 11, 2001; published electronically October 4, 2001. This research was supported by ONR/ARO Research contract DEPSCoR-DAAG55-99-1-0002.

http://www.siam.org/journals/sima/33-3/37391.html

[†] Department of Mathematics, University of Sofia, 5 James Bourchier Boulevard, 1164 Sofia, Bulgaria (nikolovg@math.bas.bg, geno@fmi.uni-sofia.bg).

$|f| \le |T_n|$ at the points of local extrema for T_n induces the inequalities $\|f^{(k)}\| \le \|T_n^{(k)}\|$ for $k = 1, \dots, n$. This suggests the following definition.

DEFINITION. *A polynomial $Q \in \pi_n$ and a mesh $\Delta = \{t_\nu\}_{\nu=0}^n$, $(1 \ge t_0 > t_1 > \cdots > t_n \ge -1)$ are said to admit the Duffin- and Schaeffer-type inequality (DS-inequality) if for every $f \in \pi_n$ the assumption $|f(t_\nu)| \le |Q(t_\nu)|$ for $\nu = 0, \dots, n$ implies $\|f'\| \le \|Q'\|$, or, more generally, $\|f^{(k)}\| \le \|Q^{(k)}\|$ for $k = 1, \dots, n$.*

Note that in our definition the comparison points $\{t_\nu\}_{\nu=0}^n$ are not necessarily assumed to be extremum points for Q.

In 1992, Shadrin [16] proposed a simple proof of the Markov inequality under the assumption (2). Based on a theorem of Shadrin, Bojanov and Nikolov [2] proved a DS-inequality for $Q = P_n^{(\lambda)}$, the ultraspherical polynomials, when the mesh Δ consists of the local extremum points of $P_n^{(\lambda)}$.

THEOREM A. *Let $Q := P_n^{(\lambda)}$ $(\lambda > -1/2)$ and $\{t_\nu\}_{\nu=0}^n$ be the zeros of $(1-x^2)Q'(x)$. If $f \in \pi_n$ satisfies*

$$|f(t_\nu)| \le |Q(t_\nu)| \quad \textit{for} \quad \nu = 0, \dots, n,$$

then

$$\|f^{(k)}\| \le \|Q^{(k)}\|$$

for all $k \in \{1, \dots, n\}$ if $\lambda \ge 0$, and for $k \ge 2$ if $\lambda \in (-1/2, 0)$. Equality is possible if and only if $f = cQ$ with $|c| = 1$.

The special case $\lambda = 0$ comes down to the classical inequality of Duffin and Schaeffer.

For some other DS-inequalities, we refer the reader to [6], [10], [11], [12], [13], [14]. In particular, the following result has been proved in [12].

THEOREM B. *If $f \in \pi_n$ satisfies $|f(\pm 1)| \le 1$ and*

$$|f(x)| \le \sqrt{1 - x^2} \quad \textit{at the zeros of } T_{n-1},$$

then

$$\|f^{(k)}\| \le \|T_n^{(k)}\| \quad \textit{for} \quad k = 1, \dots, n.$$

Moreover, equality is possible if and only $f = cT_n$ with $|c| = 1$.

Theorems A and B show that for $Q = T_n$ the DS-inequality holds at least for two choices of “check points,” namely, for those formed by the zeros of $(1 - x^2)T_n'(x)$ and by the zeros of $(1 - x^2)T_{n-1}(x)$. We naturally come to the question, What are the meshes Δ admitting the DS-inequality with $Q = T_n$? The aim of this paper is to show that for $k = 1$ each mesh $\Delta = \{t_\nu\}_{\nu=0}^n$ whose points interlace with the zeros of T_n is admissible.

THEOREM 1. *Let $\{t_\nu\}_{\nu=0}^n$ satisfy $1 \ge t_0 > \xi_1 > t_1 > \cdots > \xi_n > t_n \ge -1$, where $\{\xi_\nu\}_{\nu=1}^n$ are the zeros of T_n, i.e., $\xi_\nu = \cos((2\nu - 1)\pi/(2n))$. If $f \in \pi_n$ and*

$$|f(t_\nu)| \le |T_n(t_\nu)| \quad \textit{for} \quad \nu = 0, \dots, n,$$

then

$$\|f'\| \le n^2. \tag{3}$$

Moreover, equality in (3) is possible if and only if $f = cT_n$ with $|c| = 1$.

Note that the set of all admissible meshes Δ (i.e., such that DS-inequality holds with $Q = T_n$) cannot be substantially larger than the one described in Theorem 1. In fact, the points of any admissible mesh must separate the zeros of T_n (see section 4).

The proof of Theorem 1 relies on a pointwise inequality given by the next theorem, which was suggested to the author by Shadrin [18].

Theorem 2. *Let $Q \in \pi_n$ have n distinct zeros $\{x_\nu\}_{\nu=1}^n$, all located in $(-1, 1)$. Let $\{t_j\}_{j=0}^n$ satisfy $1 \geq t_0 > x_1 > t_1 > \cdots > x_n > t_n \geq -1$. If $f \in \pi_n$ and*

$$|f(t_j)| \leq |Q(t_j)| \quad \textit{for } j = 0, \ldots, n,$$

then for each $k \in \{1, \ldots, n\}$ and for every $x \in [-1, 1]$ there holds

$$|f^{(k)}(x)| \leq \max\{|Q^{(k)}(x)|, |Q_\nu^{(k)}(x)|,\ \nu = 1, \ldots, n\},$$

where

$$Q_\nu(x) = Q(x)\frac{1 - x_\nu x}{x - x_\nu}.$$

The paper is organized as follows. In section 2, we summarize some results from V. Markov's paper and prove Theorem 2. The proof of Theorem 1 is given in section 3. Section 4 contains some concluding remarks and points out a possible application of Theorem 1 to the estimation of the round-off error in the Lagrange differentiation formula.

2. Proof of Theorem 2.

We start with an observation from the original work of Markov [9] concerning polynomial interpolation and pointwise estimates for polynomial derivatives. We formulate it in two lemmas.

Definition. *Let $p \in \pi_n$ or $p \in \pi_{n+1}$, $q \in \pi_n$, and p, q have only real and simple zeros, say, $\{t_j\}_{j=1}^{n(+1)}$ and $\{\tau_j\}_{j=1}^n$. The zeros of p and q are said to interlace if*

$$t_1 \leq \tau_1 \leq t_2 \leq \cdots \leq t_{n-1} \leq \tau_n (\leq t_{n+1}).$$

If only strict inequalities appear above, then the zeros of p and q are said to interlace strictly.

The first of Markov's lemmas reveals a simple (and, as a matter of fact, very useful) property of the zeros of algebraic polynomials.

Lemma 3. *Let p and q be algebraic polynomials ($p \not\equiv q$), which have only real and simple zeros. If the zeros of p and q interlace, then the zeros of p' and q' interlace strictly.*

A proof of Lemma 3 can be found in [15, Lemma 2.7.1] or in [16]. Note that for polynomials of the same degree the claim of Lemma 3 can be viewed as a monotone dependence of the zeros of the derivative with respect to the zeros of the polynomial [1, p. 39].

Given a mesh $\Delta = \{t_j\}_{j=0}^n$ $(1 \geq t_0 > t_1 > \cdots > t_n \geq -1)$ and $\epsilon := \{\epsilon_j\}_{j=0}^n$ $(\epsilon_j > 0,\ j = 0, \ldots, n)$, we define the set of polynomials

$$\Omega_n(\Delta, \epsilon) := \{f \in \pi_n : |f(t_j)| \leq \epsilon_j,\ j = 0, \ldots, n\}.$$

Clearly, $\Omega_n(\Delta, \epsilon)$ is a compact set.

Define real valued polynomials $\{P_\nu\}_{\nu=0}^n = \{P_\nu(\Delta, \epsilon; \cdot)\}_{\nu=0}^n \in \Omega_n(\Delta, \epsilon)$ by

$$|P_\nu(t_j)| = \epsilon_j \quad \text{for } j, \nu = 0, \ldots, n,$$
$$P_0(t_{j-1})P_0(t_j) < 0 \quad \text{for } j = 1, \ldots, n,$$

and, for each $\nu = 1, \dots, n$,

$$P_\nu(t_{\nu-1})P_\nu(t_\nu) > 0, \quad P_\nu(t_{j-1})P_\nu(t_j) < 0 \text{ for } j \neq \nu.$$

Evidently, the above conditions determine $\{P_\nu\}_{\nu=0}^n$ uniquely up to a multiplier -1. Theorem 2 follows easily from the next lemma.

LEMMA 4. *For each $x \in [-1,1]$ and for every $k \in \{1,\dots,n\}$,*

$$\sup\{|f^{(k)}(x)| : f \in \Omega_n(\Delta,\epsilon)\} = \max\{|P_\nu^{(k)}(x)|, \nu = 0,\dots,n\}.$$

Proof. Note first that the sup is attainable since $\Omega_n(\Delta,\epsilon)$ is compact. Set $\omega(t) := (t-t_0)\cdots(t-t_n)$, $\omega_\nu(t) := \omega(t)/(t-t_\nu)$ $(\nu = 0,\dots,n)$; then for $f \in \Omega_n(\Delta,\epsilon)$ and a fixed $x \in [-1,1]$ the Lagrange interpolation formula yields

$$|f^{(k)}(x)| = \left|\sum_{j=0}^n \frac{\omega_j^{(k)}(x)}{\omega_j(t_j)} f(t_j)\right| \leq \sum_{j=0}^n \left|\frac{\omega_j^{(k)}(x)}{\omega_j(t_j)}\right| \epsilon_j. \tag{4}$$

The upper bound is attained if $|f(t_j)| = \epsilon_j$ for $j = 0,\dots,n$ and f has a suitable sign pattern at the points $\{t_j\}$. Next we show that the polynomials $\{P_\nu\}_{\nu=0}^n$ provide a complete set of appropriate sign patterns. For any pair of indices $i,j \in \{0,\dots,n\}$, $i<j$, the zeros of ω_i and ω_j interlace (though not strictly); therefore, in view of Lemma 3, the zeros $\{\gamma_{i,\mu}\}_{\mu=1}^{n-k}$ of $\omega_i^{(k)}$ and the zeros $\{\gamma_{j,\mu}\}_{\mu=1}^{n-k}$ of $\omega_j^{(k)}$ interlace strictly. Furthermore, since the zeros of ω_i are less than or equal to the corresponding zeros of ω_j, we have the following arrangement:

$$\gamma_{0,n-k} < \cdots < \gamma_{n,n-k} < \gamma_{0,n-k-1} < \cdots < \gamma_{n,n-k-1} < \cdots < \gamma_{0,1} < \cdots < \gamma_{n,1}.$$

Since $\omega_{j-1}(t_{j-1})\omega_j(t_j) < 0$ for $j = 1,\dots,n$, the above inequalities show that for $x \in [-1,1] \setminus \{\gamma_{\nu,j}\}_{\nu=0,j=1}^{n\;\;n-k}$, the quantities $\{\omega_j^{(k)}(x)/\omega_j(t_j)\}_{j=0}^n$ either change their signs alternatively if

$$x \in I_{n,k}^0, \quad I_{n,k}^0 = I_{n,k}^0(\Delta) := [-1,\gamma_{0,n-k}) \bigcup_{j=n-k}^{1} (\gamma_{n,j},\gamma_{0,j-1}) \bigcup (\gamma_{n,1},1]$$

or change signs alternatively with only one exception: $\frac{\omega_{\nu-1}^{(k)}(x)}{\omega_{\nu-1}(t_{\nu-1})}\frac{\omega_\nu^{(k)}(x)}{\omega_\nu(t_\nu)} > 0$ for some $\nu \in \{1,\dots,n\}$. The latter situation occurs when $x \in I_{n,k}^\nu$, where

$$I_{n,k}^\nu = I_{n,k}^\nu(\Delta) := \bigcup_{j=1}^{n-k} (\gamma_{\nu-1,j},\gamma_{\nu,j}).$$

Correspondingly, if $x \in I_{n,k}^\nu$ for some $\nu \in \{0,\dots,n\}$, then (4) holds with equality sign for $f = P_\nu$. If $x = \gamma_{\nu,j}$, then $\omega_\nu^{(k)}(x) = 0$, and equality in (4) holds for $f = P_\nu$ as well as for any $f \in \pi_n$ which coincides with P_ν at the points $\{t_j : j \neq \nu\}$.

Thus, in (4), equality holds for $f = P_\nu$ if $x \in \overline{I}_{n,k}^\nu$ $(\nu = 0,\dots,n)$, and since $\cup_{\nu=0}^n \overline{I}_{n,k}^\nu = [-1,1]$, the proof of Lemma 4 is complete. □

Remark. It follows from the proof of Lemma 4 that if for some $f \in \Omega_n(\Delta,\epsilon)$ we have equality in (4) for some $x \in I_{n,k}^\nu$ $(\nu \in \{0,\dots,n\})$, then necessarily $f = cP_\nu$, where c is a constant with $|c| = 1$. Thus, for $x \in [-1,1] \setminus \{\gamma_{\nu,j}\}_{\nu=0,j=1}^{n\;\;n-k}$, any extremal polynomial in Lemma 4 is of the form $f = cP_\nu$, where $\nu \in \{0,\dots,n\}$ and $|c| = 1$.

Proof of Theorem 2. Set $\epsilon_j := |Q(t_j)|$, $j = 0, \dots, n$, and define polynomials $\{P_\nu\}_{\nu=0}^n$ as above. Based on the interlacing assumption, we conclude that $P_0 = Q$ or $P_0 = -Q$, while for $\nu = 1, \dots, n$ the sign patterns of P_ν and Q_ν coincide. Moreover, we have

$$|Q_\nu(t_j)| = \epsilon_j \frac{1 - x_\nu t_j}{|t_j - x_\nu|} \geq \epsilon_j \text{ for } j = 0, \dots, n \quad \text{and} \quad \nu = 1, \dots, n.$$

In the proof of Lemma 4, we deduced that for any $f \in \Omega_n(\Delta, \epsilon)$

$$|f^{(k)}(x)| \leq |P_\nu^{(k)}(x)| \text{ if } x \in \overline{I}_{n,k}^{\nu}, \quad \nu = 0, \dots, n. \tag{5}$$

For $\nu = 0$ (5) reads as $|f^{(k)}(x)| \leq |Q^{(k)}(x)|$, while for $x \in \overline{I}_{n,k}^{\nu}$ ($\nu \in \{1, \dots, n\}$) we have

$$|P_\nu^{(k)}(x)| = \sum_{j=0}^{n} \left| \frac{\omega_j^{(k)}(x)}{\omega_j(t_j)} \right| \epsilon_j \leq \sum_{j=0}^{n} \left| \frac{\omega_j^{(k)}(x)}{\omega_j(t_j)} \right| |Q_\nu(t_j)| = |Q_\nu^{(k)}(x)|.$$

(For the last equality we used that P_ν and Q_ν have the same sign pattern.) The claim of Theorem 2 now follows from Lemma 4. □

As an immediate consequence of Theorem 2 we get the following corollary.

Corollary 5. *If, in addition to the assumptions of Theorem* 2, *for a* $k \in \{1, \dots, n\}$

$$\max_{1 \leq \nu \leq n} \|Q_\nu^{(k)}\| \leq \|Q^{(k)}\|,$$

then

$$\|f^{(k)}\| \leq \|Q^{(k)}\|.$$

3. Proof of Theorem 1.

The proof of Theorem 1 follows from Corollary 5, applied to $Q = T_n$ with $k = 1$. The application of Corollary 5 is possible because of the following lemma.

Lemma 6. *Let the polynomials* $\{P_\nu\}_{\nu=1}^n$ *be defined by*

$$P_\nu(x) := T_n(x) \frac{1 - \xi_\nu x}{x - \xi_\nu}.$$

Then, for $n \geq 2$,

$$\|P_\nu'\| < n^2 \quad (\nu = 1, \dots, n). \tag{6}$$

For $n = 2, 3$ the validity of (6) is verified directly; therefore, we assume in what follows that $n \geq 4$. The proof of Lemma 6 goes through a number of lemmas.

Lemma 7. *For every* $x \in [-1, 1]$ *and for* $\nu = 1, \dots, n$

$$|P_\nu'(x)| \leq R_\nu(x),$$

where

$$R_\nu(x) = \left[\frac{(1 - \xi_\nu^2)^2}{(x - \xi_\nu)^4} + \frac{n^2 (1 - \xi_\nu x)^2}{(1 - x^2)(x - \xi_\nu)^2} \right]^{1/2}.$$

Proof. The result is immediate from

$$P_\nu'(x) = T_n'(x)\frac{1-\xi_\nu x}{x-\xi_\nu} - T_n(x)\frac{1-\xi_\nu^2}{(x-\xi_\nu)^2}, \tag{7}$$

the identity $[T_n(x)]^2 + (1-x^2)[T_n'(x)]^2/n^2 = 1$, and Cauchy's inequality. □

LEMMA 8. *$R_\nu(x)$ is a strictly convex function on each of the intervals $(-1,\xi_\nu)$ and $(\xi_\nu,1)$.*

Proof. We suppress the index ν, writing

$$R(x) = \left[\frac{(1-\xi^2)^2}{(x-\xi)^4} + \frac{n^2(1-\xi x)^2}{(1-x^2)(x-\xi)^2}\right]^{1/2} =: (g_1^2(x) + g_2^2(x))^{1/2},$$

where

$$g_1(x) := \frac{1-\xi^2}{(x-\xi)^2},\ g_2(x) := \frac{n(1-\xi x)}{(1-x^2)^{1/2}(x-\xi)}.$$

Since

$$R'' = \frac{(g_1g_2' - g_1'g_2)^2 + R^2(g_1g_1'' + g_2g_2'')}{R^3},$$

the lemma will be proved if we show that $g_1(x)g_1''(x)$ and $g_2(x)g_2''(x)$ are positive in $(-1,\xi)$ and in $(\xi,1)$. This is easily seen for the first term, while for the second term a short calculation yields

$$\begin{aligned}&\frac{(x-\xi)^4(1-x^2)^3}{n^2}g_2(x)g_2''(x)\\ &\quad = 2(1-\xi^2)(1-x^2)^2 - 2x(x-\xi)(1-\xi^2)(1-x^2) + (1-\xi x)(x-\xi)^2(2x^2+1).\end{aligned}$$

The positivity of the right-hand side is easily verified with the help of the inequality

$$\begin{aligned}&2(1-\xi^2)(1-x^2)^2 + (1-\xi x)(x-\xi)^2(2x^2+1)\\ &\quad \ge 2(1-x^2)|x-\xi|[2(1-\xi^2)(1-\xi x)(2x^2+1)]^{1/2}.\end{aligned}$$ □

We now examine the polynomials $\{P_\nu\}_{\nu=1}^n$. Due to symmetry, we may (and shall) consider only half of them, say, those with indices $1 \le \nu \le [(n+1)/2]$. Recall that the zeros of P_ν coincide with the zeros $\{\xi_j\}_{j=1}^n$ of T_n with the exception of ξ_ν which is replaced by $1/\xi_\nu$. (In the case where n is odd and $\nu = (n+1)/2$, $1/\xi_\nu$ is interpreted as a zero at ∞.) With this last convention, we observe that for $1 \le \nu \le [(n+1)/2]$ the zeros of P_ν are located to the right with respect to the zeros $\{\xi_i\}$ of T_n and interlace with them. In view of Lemma 3, the same relation holds between the zeros of the derivatives of P_ν and T_n. We are interested in the behavior of $P_\nu'(x)$ and, in particular, its critical points. To this end, we shall exploit (7) and the explicit form of P_ν'',

$$P_\nu''(x) = T_n''(x)\frac{1-\xi_\nu x}{x-\xi_\nu} - 2T_n'(x)\frac{1-\xi_\nu^2}{(x-\xi_\nu)^2} + 2T_n(x)\frac{1-\xi_\nu^2}{(x-\xi_\nu)^3}. \tag{8}$$

In the proof of the next lemmas we shall use the differential equation

$$(1-x^2)T_n''(x) - xT_n'(x) + n^2T_n(x) = 0 \tag{9}$$

as well as the following simple facts:

$$\{n\sin(\alpha/n)\}_{n=1}^{\infty}\nearrow\alpha, \tag{10}$$

$$\cot\alpha\le\frac{1}{\alpha}, \tag{11}$$

where $0<\alpha\le\pi/2$.

Lemma 9. *The polynomials* P_ν' $(\nu=1,\dots,[(n+1)/2])$ *satisfy the following:*

(i) *if* $2\le\nu<\frac{n+1}{2}$, *then* P_ν' *has exactly one local extremum to the right of* 1;

(ii) P_ν' *has exactly one local extremum in* $(\xi_{\nu+1},\eta_\nu)$;

(iii) P_ν' *is strictly monotone in* $[\eta_\nu,\eta_{\nu-1}]$;

(iv) P_ν' *is strictly monotone in* $[-1,\eta_{n-1}]$ *and in* $[\eta_1,1]$.

Proof. The first claim in (iv) follows trivially since, as was already mentioned, the zeros of P_ν are located to the right with respect to $\{\xi_j\}_{j=1}^n$. In view of Lemma 3, the same is true for the zeros of P_ν'' and T_n''. Since the leftmost zero of T_n'' is located to the right of η_{n-1}, so is the smallest zero of P_ν''.

Substituting $x=1$ in (8), we get

$$P_\nu''(1)=\frac{n^2(n^2-1)}{3}-2n^2\cot^2\frac{(2\nu-1)\pi}{4n}+\frac{\cot^2\frac{(2\nu-1)\pi}{4n}}{\sin^2\frac{(2\nu-1)\pi}{4n}}.$$

With the help of (10) and (for $\nu=2$) (11), it is easy to see that $P_\nu''(1)>0$ for $2\le\nu\le[(n+1)/2]$. Since P_ν' has a negative leading coefficient and at most one critical point to the right of $x=1$, this proves part (i) of the lemma.

Now we find the sign of P_ν'' at the points $\xi_{\nu+1}$, η_ν, and $\eta_{\nu-1}$. First, we shall show that

$$\operatorname{sign}\{P_\nu''(\xi_{\nu+1})\}=(-1)^{\nu+1}. \tag{12}$$

Putting $x=\xi_{\nu+1}$ in (8) and using that $T_n''(\xi_{\nu+1})=\xi_{\nu+1}T_n'(\xi_{\nu+1})/(1-\xi_{\nu+1}^2)$ and $\operatorname{sign}\{T_n'(\xi_{\nu+1})\}=(-1)^\nu$, we get

$$\operatorname{sign}\{P_\nu''(\xi_{\nu+1})\}=(-1)^{\nu+1}\operatorname{sign}\{2(1-\xi_\nu^2)(1-\xi_{\nu+1}^2)+\xi_{\nu+1}(\xi_\nu-\xi_{\nu+1})(1-\xi_\nu\xi_{\nu+1})\}.$$

Now (12) is obvious if $\xi_{\nu+1}\ge0$. The only possibility where $\xi_{\nu+1}<0$ is $\nu=m$ and $n=2m$ or $n=2m-1$. An easy calculation shows that for $n\ge4$ (12) is true in this case, too.

Next we prove both (ii) and (iii) by showing that

$$\operatorname{sign}\{P_\nu''(\eta_\mu)\}=(-1)^\nu\quad\text{for } \mu=\nu,\ \nu-1,\ \mu\ne0. \tag{13}$$

Using (8) and (9), we obtain

$$P_\nu''(\eta_\mu)=\frac{T_n(\eta_\mu)}{(\xi_\nu-\eta_\mu)^3(1-\eta_\mu^2)}[n^2(1-\xi_\nu\eta_\mu)(\xi_\nu-\eta_\mu)^2-2(1-\xi_\nu^2)(1-\eta_\mu^2)]. \tag{14}$$

Since $\operatorname{sign}\{T_n(\eta_\mu)\}=(-1)^\mu$, it suffices to prove that the term in the square brackets is positive. Using the inequality $(1-\xi_\nu^2)(1-\eta_\mu^2)<(1-\xi_\nu\eta_\mu)^2$, we obtain

$$n^2(1-\xi_\nu\eta_\mu)(\xi_\nu-\eta_\mu)^2-2(1-\xi_\nu^2)(1-\eta_\mu^2)>(1-\xi_\nu\eta_\mu)[n^2(\xi_\nu-\eta_\mu)^2-2(1-\xi_\nu\eta_\mu)].$$

After simple manipulations, using the trigonometric representation of ξ_ν and η_μ, we find that the inequality $n^2(\xi_\nu - \eta_\mu)^2 - 2(1 - \xi_\nu\eta_\mu) \geq 0$ is equivalent to

$$\frac{1}{n^2 \sin^2 \frac{\pi}{4n}} + \frac{1}{n^2 \sin^2 \frac{(2\nu+2\mu-1)\pi}{4n}} \leq 2.$$

This last inequality will hold for all $\nu \in \{1, \ldots, [(n+1)/2]\}$ and $\mu = \nu$, $\nu - 1$, $(\mu \neq 0)$ if it is true for $\nu = \mu = 1$, i.e., if

$$\frac{1}{n^2 \sin^2 \frac{\pi}{4n}} + \frac{1}{n^2 \sin^2 \frac{3\pi}{4n}} \leq 2.$$

Since the left-hand side is a decreasing function of n (see (10)), and for $n = 3$ it is $(\sin^{-2}(\pi/12) + 2)/9 = (4\sqrt{3} + 10)/9 < 2$, (13) is proved. Now we conclude from (12) and (13) (with $\mu = \nu$) that P''_ν has a zero in $(\xi_{\nu+1}, \eta_\nu)$ for $\nu = 1, \ldots, [(n+1)/2]$. In addition, (13) implies that this zero is unique, and no zeros of P''_ν exist in $[\eta_\nu, \eta_{\nu-1}]$ $(\nu \geq 2)$; otherwise, there would be at least three zeros in $(\xi_{\nu+1}, \xi_{\nu-1})$, which is a contradiction. For the same reason, P''_1 has a simple zero in (ξ_2, η_1), and no zeros of P''_1 exist in $[\eta_1, 1]$. This is exactly the claim of (iii) for $\nu = 1$ and of the second part of (iv) for $\nu = 1$.

To prove the second part of (iv) for $2 \leq \nu < (n+1)/2$, we shall show that

$$P''_\nu(\eta_1) > 0. \tag{15}$$

Having established (15), the second part of (iv) will follow immediately. Indeed, we found in the beginning of this proof that $P''_\nu(1) > 0$ for $2 \leq \nu < (n+1)/2$, and if P'_ν was not monotone in $[\eta_1, 1]$, then P''_ν would have at least three zeros (two zeros if $\nu = (n+1)/2$)) to the right of η_1, which is impossible. The proof of (15) goes along the lines of the proof of (13). Equation (14) with $\mu = 1$ shows that (15) follows if

$$n^2(1 - \xi_\nu\eta_1)(\xi_\nu - \eta_1)^2 - 2(1 - \xi_\nu^2)(1 - \eta_1^2) > 0$$

or, in view of $(1 - \xi_\nu^2)(1 - \eta_1^2) \leq (1 - \xi_\nu\eta_1)^2$, if

$$n^2(\xi_\nu - \eta_1)^2 - 2(1 - \xi_\nu\eta_1) > 0.$$

The latter inequality is equivalent to the inequality

$$\frac{1}{n^2 \sin^2 \frac{(2\nu-3)\pi}{4n}} + \frac{1}{n^2 \sin^2 \frac{(2\nu+1)\pi}{4n}} \leq 2,$$

whose validity is easily verified with the help of (10). Lemma 9 is proved. □

LEMMA 10. *The following estimates for* $\|P'_\nu\|$ *hold true:*

(i) *for* $\nu = 1, 2$,

$$\|P'_\nu\| \leq \max\{|P'_\nu(-1)|, |P'_\nu(1)|, R_\nu(\eta_{n-1}), R_\nu(\eta_\nu)\};$$

(ii) *for* $\nu = 3, \ldots, [(n+1)/2]$,

$$\|P'_\nu\| \leq \max\{|P'_\nu(-1)|, |P'_\nu(1)|, R_\nu(\eta_{n-1}), R_\nu(\eta_\nu), R_\nu(\eta_{\nu-1}), R_\nu(\eta_1)\}.$$

Proof. According to Lemma 9, P'_1 is monotone in $[-1, \eta_{n-1}]$ and $[\eta_1, 1]$; therefore, on these intervals,

$$|P'_1(x)| \leq \max\{|P'_1(-1)|, |P'_1(\eta_{n-1})|, |P'_1(\eta_1)|, |P'_1(1)|\}.$$

On the complementary interval $[\eta_{n-1},\eta_1]$, we have $|P_1'(x)|\le R_1(x)$ (Lemma 7), and since R_1 is convex there (Lemma 8), it follows that $R_1(x)\le\max\{R_1(\eta_{n-1}),R_1(\eta_1)\}$ for $x\in[\eta_{n-1},\eta_1]$. This proves (i) for $\nu=1$.

The proof of (i) for $\nu=2$ relies on the observation that, by Lemma 9, P_2' is monotone in $[-1,\eta_{n-1}]$ and in $[\eta_2,1]$, while $|P_2'(x)|\le\max\{R_2(\eta_{n-1}),R_2(\eta_2)\}$ in $[\eta_{n-1},\eta_2]$, by virtue of Lemmas 7 and 8.

Part (ii) can be proved in the same fashion, exploiting the monotonicity of P_ν' on the intervals $[-1,\eta_{n-1}]$, $[\eta_\nu,\eta_{\nu-1}]$, and $[\eta_1,1]$ and the convexity of R_ν on $[\eta_{n-1},\eta_\nu]$ and $[\eta_{\nu-1},\eta_1]$. We leave the details to the reader. □

Our last lemma estimates the quantities appearing in Lemma 10.

Lemma 11. *The following inequalities hold true:*

(i) $|P_\nu'(\pm1)|<n^2$ $(\nu=1,\dots,[(n+1)/2])$;

(ii) $R_\nu(\eta_1)<n^2$ $(\nu=1,3,4,\dots,[(n+1)/2])$;

(iii) $R_\nu(\eta_\nu)<n^2$ $(\nu=1,\dots,[(n+1)/2])$;

(iv) $R_\nu(\eta_{\nu-1})<n^2$ $(\nu=3,\dots,[(n+1)/2])$;

(v) $R_\nu(\eta_{n-1})<n^2$ $(\nu=1,\dots,[(n+1)/2])$.

Proof. Substituting $x=\pm1$ in (7), we get

$$P_\nu'(1)=n^2-\cot^2\frac{(2\nu-1)\pi}{4n},\quad |P_\nu'(-1)|=n^2-\tan^2\frac{(2\nu-1)\pi}{4n}.$$

Then (10) and $0<(2\nu-1)\pi/(2n)\le\pi/4$ show the validity of slightly sharper inequalities than (i), namely,

$$n^2-1\le|P_\nu'(-1)|<n^2-\frac{\pi}{4n}$$

and

$$(1-16/\pi^2)n^2<P_\nu'(1)<n^2-1.$$

Now we prove (ii). A short calculation yields

$$R_\nu(\eta_1)=\left[\frac{(1-\xi_\nu^2)^2}{(\eta_1-\xi_\nu)^4}+\frac{n^2(1-\xi_\nu\eta_1)^2}{(1-\eta_1^2)(\eta_1-\xi_\nu)^2}\right]^{1/2}=:\{[A(\nu)]^2+[B(\nu)]^4\}^{1/2},$$

where

$$A(\nu)=\frac{n}{2}\left|2\cot\frac{\pi}{n}+\cot\frac{(2\nu-3)\pi}{4n}-\cot\frac{(2\nu+1)\pi}{4n}\right|,$$
$$B(\nu)=\frac{1}{2}\left|\cot\frac{(2\nu-3)\pi}{4n}+\cot\frac{(2\nu+1)\pi}{4n}\right|.$$

Assume first that $3\le\nu\le[(n+1)/2]$; then it is easy to see that $A(\nu)\le A(3)$ and $B(\nu)\le B(3)$. We use (11) to obtain

$$\begin{aligned}B(3)&=\frac12\left[\cot\frac{3\pi}{4n}+\cot\frac{7\pi}{4n}\right]<\frac{20n}{21\pi},\\ A(3)&=\frac{n}{2}\left[\cot\frac{3\pi}{4n}+2\cot\frac{\pi}{n}-\cot\frac{7\pi}{4n}\right]\\ &<\frac{n}{2}\left[\cot\frac{3\pi}{4n}+2\cot\frac{\pi}{n}\right]\\ &<\frac{5n^2}{3\pi}.\end{aligned}$$

Therefore, for $3 \leq \nu \leq [(n+1)/2]$,

$$R_\nu(\eta_1) < \left[\left(\frac{5n^2}{3\pi}\right)^2 + \left(\frac{20n}{21\pi}\right)^4\right]^{1/2} < 0.54n^2 < n^2.$$

Similarly, for $\nu = 1$, we find

$$\begin{aligned} A(1) &= \frac{n}{2}\left[\cot\frac{\pi}{4n} - 2\cot\frac{\pi}{n} + \cot\frac{3\pi}{4n}\right] < \frac{n}{2}\left[\cot\frac{\pi}{4n} + \cot\frac{3\pi}{4n}\right] < \frac{8n^2}{3\pi}, \\ B(1) &= \frac{1}{2}\left[\cot\frac{\pi}{4n} - \cot\frac{3\pi}{4n}\right] < \frac{1}{2}\cot\frac{\pi}{4n} < \frac{2n}{\pi}. \end{aligned}$$

Hence

$$R_1(\eta_1) < \left[\left(\frac{8n^2}{3\pi}\right)^2 + \left(\frac{2n}{\pi}\right)^4\right]^{1/2} < 0.95n^2 < n^2.$$

Thus (ii) is proved.

Next we prove (iii). For $1 \leq \nu \leq [(n+1)/2]$, we have

$$R_\nu(\eta_\nu) = \left[\frac{(1-\xi_\nu^2)^2}{(\xi_\nu - \eta_\nu)^4} + \frac{n^2(1-\xi_\nu\eta_\nu)^2}{(1-\eta_\nu^2)(\xi_\nu - \eta_\nu)^2}\right]^{1/2} =: \{[C(\nu)]^2 + [D(\nu)]^4\}^{1/2},$$

where

$$\begin{aligned} C(\nu) &= \frac{n}{2}\left[\cot\frac{\pi}{4n} + \cot\frac{(4\nu-1)\pi}{4n} - 2\cot\frac{\nu\pi}{n}\right], \\ D(\nu) &:= \frac{1}{2}\left[\cot\frac{\pi}{4n} - \cot\frac{(4\nu-1)\pi}{4n}\right]. \end{aligned}$$

Unlike the situation with $A(\nu)$ and $B(\nu)$, we observe that $C(\nu)$ and $D(\nu)$ increase with ν, and for $n \geq 3$

$$\begin{aligned} D(\nu) \leq D((n+1)/2) &= \frac{n}{n\sin\frac{\pi}{2n}} \leq \frac{2n}{3}, \\ C(\nu) \leq C((n+1)/2) &= \frac{n}{2}\left[\cot\frac{\pi}{4n} + 2\tan\frac{\pi}{2n} - \tan\frac{\pi}{4n}\right] \\ &= \frac{n}{\sin\frac{\pi}{2n}} + n\left[\tan\frac{\pi}{2n} - \tan\frac{\pi}{4n}\right] \\ &< \frac{n^2}{n\sin\frac{\pi}{2n}} + \frac{\pi}{4}\frac{1}{\cos^2\frac{\pi}{2n}} \\ &\leq \frac{1}{3}(2n^2 + \pi). \end{aligned}$$

With this (iii) is proved since

$$R_\nu(\eta_\nu) < n^2\left[\left(\frac{2}{3} + \frac{\pi}{3n^2}\right)^2 + \left(\frac{2}{3}\right)^4\right]^{1/2} < 0.91n^2 < n^2.$$

The same arguments as above lead to the proof of (iv): $R_\nu(\eta_{\nu-1}) = [(\tilde{C}(\nu))^2 + (\tilde{D}(\nu))^4]^{1/2}$, where

$$\begin{aligned}\tilde{C}(\nu) &= \frac{n}{2}\left[\cot\frac{\pi}{4n} + 2\cot\frac{(\nu-1)\pi}{n} - \cot\frac{(4\nu-3)\pi}{4n}\right],\\ \tilde{D}(\nu) &= \frac{1}{2}\left[\cot\frac{\pi}{4n} + \cot\frac{(4\nu-3)\pi}{4n}\right].\end{aligned}$$

Observing that $\tilde{C}(\nu)$ and $\tilde{D}(\nu)$ decrease with ν, for $3 \le \nu \le [(n+1)/2]$ we find the estimates

$$\begin{aligned}\tilde{D}(\nu) \le \tilde{D}(3) &= \frac{1}{2}\left[\cot\frac{\pi}{4n} + \cot\frac{9\pi}{4n}\right] < \frac{20n}{9\pi},\\ \tilde{C}(\nu) \le \tilde{C}(3) &= \frac{n}{2}\left[\cot\frac{\pi}{4n} + 2\cot\frac{2\pi}{n} - \cot\frac{9\pi}{4n}\right]\\ &< \frac{n}{2}\left[\cot\frac{\pi}{4n} + \cot\frac{7\pi}{4n}\right]\\ &< \frac{16n^2}{7\pi},\end{aligned}$$

and hence

$$R_\nu(\eta_{\nu-1}) < \left[\left(\frac{16n^2}{7\pi}\right)^2 + \left(\frac{20n}{9\pi}\right)^4\right]^{1/2} < 0.89n^2 < n^2.$$

Finally, (v) can be proved in the same way as (i)–(iv). Alternatively, one can use the inequality

$$\frac{1-\xi\eta}{|\xi-\eta|} \ge \frac{1+\xi\eta}{\xi+\eta} \quad (0 \le \xi, \eta < 1,\ \xi \ne \eta)$$

to compare pairwise $A(\nu)$ and $B(\nu)$ with the corresponding terms in $R_\nu(\eta_{n-1}) = R_\nu(-\eta_1)$. The result is $R_\nu(\eta_{n-1}) \le R_\nu(\eta_1) < n^2$. We omit the details. □

Proof of Lemma 6. The inequality follows from Lemmas 10 and 11. □

Proof of Theorem 1. Inequality (3) follows immediately from Corollary 5 and Lemma 6. It remains to clarify in which cases an equality is possible. Let $\Delta = \{t_j\}_{j=0}^n$ be a fixed mesh satisfying the assumptions of Theorem 1. Let $\epsilon = (\epsilon_0, \dots, \epsilon_n) =: (|T_n(t_0)|, \dots, |T_n(t_n)|)$, and let the polynomials $P_0 = T_n,\ P_\nu\ (\nu = 1, \dots, n)$ be defined as in section 2. Suppose that $f \in \Omega(\Delta, \epsilon)$ is an extremal polynomial, i.e., $\|f'\| = n^2$. According to Lemma 6 and the remark following the proof of Lemma 4, for $x \in \cup_{\nu=1}^n \overline{I}_{n,1}^\nu$ there holds

$$|f'(x)| \le \max_{1\le\nu\le n} \|P'_\nu\| < n^2;$$

therefore, $\|f'\|$ is attained for $x \in I_{n,1}^0$. However, when $x \in I_{n,1}^0$ we have

$$|f'(x)| \le |P'_0(x)| = |T'_n(x)| \le T'_n(1) = n^2,$$

and equality holds only for $x = \pm 1$ and $f = cT_n$ with $|c| = 1$. Theorem 1 is proved. □

4. Concluding remarks.

1. The requirement in Theorem 1 that the points $\Delta = \{t_j\}_{j=0}^n$ interlace strictly with the zeros of T_n was imposed only in order to avoid unimportant complications in the proof. Actually, Theorem 1 is valid under the weaker assumption that $\{t_j\}_{j=0}^n$ interlace with $\{\xi_j\}_{j=1}^n$. If a comparison point t_j coincides with a zero of T_n, then the polynomials from the corresponding class $\Omega_n(\Delta, \epsilon)$ must vanish at that point. In the case when all $\{\xi_\nu\}_{\nu=1}^n$ belong to Δ, Theorem 1 holds trivially since in that case $\Omega_n(\Delta, \epsilon) = \{cT_n(x) : |c| \le 1\}$.

2. So far, we cannot extend Theorem 1 to higher order derivatives, i.e., to prove $\|f^{(k)}\| \le \|T_n^{(k)}\|$ for all $k \ge 2$. However, it should be pointed out that this inequality holds true for $k = n-1$ and for $k = n$. This is easily seen from the proof of Lemma 4: for any polynomial $f \in \Omega_n(\Delta, \epsilon)$ and for $k = n-1, n$ we have $\|f^{(k)}\| = |f^{(k)}(-1)|$ or $\|f^{(k)}\| = |f^{(k)}(1)|$, and for $x = \pm 1$ the extremal polynomials in Lemma 4 are of the form $cP_0 = \pm cT_n$, $|c| = 1$.

3. According to Lemma 4, a necessary condition for a mesh $\Delta = \{t_j\}_{j=0}^n$ to admit DS-inequality with an extremal polynomial $Q = T_n$ is for the sign pattern of $(T_n(t_0), \ldots, T_n(t_n))$ to coincide (up to a factor -1) with the sign pattern of some of the polynomials $\{P_\nu\}_{\nu=0}^n$. Theorem 1 asserts DS-inequality for all meshes Δ having the sign structure of P_0. One may think that DS-inequality also holds for any other mesh $\Delta = \{t_j\}_{j=0}^n$ for which the sign pattern of $(T_n(t_0), \ldots, T_n(t_n))$ coincides with the sign pattern of some P_ν, $\nu \in \{1, \ldots, n\}$. However, the example below shows that this is not true in general.

Let $t_j = \eta_{j+1}$ for $j = 0, 1, \ldots, n-2$, $t_n = \eta_n$, and $t_{n-1} = \zeta$, where $\zeta \in (-1, \xi_n)$. Define the polynomial

$$q(x) = \begin{cases} T_n(x) & \text{for } x = t_j, \quad j = 0, \ldots, n-2, n, \\ \\ -T_n(x) & \text{for } x = t_{n-1}. \end{cases}$$

Clearly, q has the same sign structure as P_{n-1}, and $|q(t_j)| = |T_n(t_j)|$ $(j = 0, \ldots, n)$. The explicit form of q is

$$q(x) = T_n(x) + a(1+x)T_n'(x), \quad \text{where } a = -2T_n(\zeta)/((1+\zeta)T_n'(\zeta)) > 0,$$

and for $k = 1, \ldots, n$ we have

$$\|q^{(k)}\| \ge q^{(k)}(1) > T_n^{(k)}(1) = \|T_n^{(k)}\|.$$

4. As was mentioned in [11, p. 174], inequalities of Duffin–Schaeffer type may be viewed as exact estimates for the round-off error in Lagrange differentiation formulas. We describe below briefly a possible application of the result of Theorem 1.

Let $\Delta = \{t_j\}_{j=0}^n$ be a mesh whose points interlace strictly with the zeros of T_n. Suppose that inaccurate data $\{\tilde{f}(t_j)\}_{j=0}^n$ for a function $f \in C^{n+1}[-1, 1]$ is given, where

$$|f(t_j) - \tilde{f}(t_j)| \le \delta_j \quad (j = 0, \ldots, n).$$

If $f'(x) \approx L_n'(\tilde{f}; x)$ is the Lagrange differentiation formula based on this information, then for the error $R(f; x) := f'(x) - L_n'(\tilde{f}; x)$ there holds

$$R(f; x) = R^{\text{round}}(f; x) + R^{\text{trunc}}(f; x)$$

with $R^{\text{round}}(f;x) = L_n'(\tilde{f} - f;x)$ being the error caused by inaccuracy of the data and $R^{\text{trunc}}(f;x)$ being the error caused by the fact that f is not necessarily a polynomial (truncation error). We have the estimate

$$\|R(f;\cdot)\| \le \|R^{\text{round}}(f;\cdot)\| + \|R^{\text{trunc}}(f;\cdot)\|.$$

The exact bound for the truncation error in the Lagrange differentiation formula in the general case has been obtained by Shadrin [17] (in our case $\|R^{\text{trunc}}(f;\cdot)\| \le \|f^{(n+1)}\|\|\omega'\|/(n+1)!$). For the round-off error, Theorem 1 provides the following exact upper bound:

$$\|R^{\text{round}}(f;\cdot)\| \le Mn^2, \quad \text{where} \quad M = \max_{0\le j\le n} \frac{\delta_j}{|T_n(t_j)|}.$$

This upper bound is attained when $\delta_j/|T_n(t_j)| = M$ for $j = 0,\dots,n$.

REFERENCES

[1] B. D. Bojanov, *An inequality of Duffin and Schaeffer type*, East J. Approx., 1 (1995), pp. 37–46.

[2] B. D. Bojanov and G. P. Nikolov, *Duffin and Schaeffer type inequality for ultraspherical polynomials*, J. Approx. Theory, 84 (1996), pp. 129–138.

[3] P. Borwein and T. Erdélyi, *Polynomials and Polynomial Inequalities*, Springer, Berlin, 1995.

[4] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, Springer, Berlin, 1993.

[5] R. J. Duffin and A. C. Schaeffer, *A refinement of an inequality of brothers Markoff*, Trans. Amer. Math. Soc., 50 (1941), pp. 517–528.

[6] C. Frappier, *On the inequalities of Bernstein–Markoff for an interval*, J. Anal. Math., 43 (1983–1984), pp. 12–25.

[7] G. G. Lorentz, M. V. Golitschek, and Y. Makovoz, *Constructive Approximation. Advanced Problems*, Springer, Berlin, 1996.

[8] A. A. Markov, *On a question of D. I. Mendeleev*, Zap. Petersburg. Akad. Nauk, 62 (1889), pp. 1–24 (in Russian).

[9] V. A. Markov, *On functions least deviated from zero in a given interval*, St. Petersburg, 1892 (in Russian); German translation: *Über Polynome, die in einem gegeben Intervalle möglichst wenig von Null abweichen*, Math. Ann., 77 (1916), pp. 213–258.

[10] L. B. Milev and G. P. Nikolov, *On the inequality of I. Schur*, J. Math. Anal. Appl., 16 (1997), pp. 421–437.

[11] G. P. Nikolov, *On certain Duffin and Schaeffer type inequalities*, J. Approx. Theory, 93 (1998), pp. 157–176.

[12] G. P. Nikolov, *An inequality for polynomials with elliptic majorant*, J. Inequal. Appl., 4 (1999), pp. 315–325.

[13] Q. I. Rahman and G. Schmeisser, *Markov–Duffin–Schaeffer inequality for polynomials with a circular majorant*, Trans. Amer. Math. Soc., 310 (1988), pp. 693–702.

[14] Q. I. Rahman and A. Q. Watt, *Polynomials with a parabolic majorant and the Duffin–Schaeffer inequality*, J. Approx. Theory 69 (1992), pp. 338–354.

[15] T. J. Rivlin, *The Chebyshev Polynomials*, Wiley, New York, 1974.

[16] A. Yu. Shadrin, *Interpolation with Lagrange polynomials: A simple proof of Markov inequality and some of its generalizations*, Approx. Theory Appl. (N.S.), 8 (1992), pp. 51–61.

[17] A. Yu. Shadrin, *Error bound for Lagrange interpolation*, J. Approx. Theory, 80 (1995), pp. 25–49.

[18] A. Yu. Shadrin, *Private communication*, Cambridge University, Department of Applied Mathematics and Theoretical Physics, Cambridge, UK, 1996.

SIAM J. MATH. ANAL.
Vol. 33, No. 3, pp. 699–717

ON THE EQUIVALENCE OF VISCOSITY SOLUTIONS AND WEAK SOLUTIONS FOR A QUASI-LINEAR EQUATION*

PETRI JUUTINEN[†], PETER LINDQVIST[‡], AND JUAN J. MANFREDI[§]

Abstract. We discuss and compare various notions of weak solution for the p-Laplace equation

$$-\mathrm{div}(|\nabla u|^{p-2}\nabla u) = 0$$

and its parabolic counterpart

$$u_t - \mathrm{div}(|\nabla u|^{p-2}\nabla u) = 0.$$

In addition to the usual Sobolev weak solutions based on integration by parts, we consider the p-superharmonic (or p-superparabolic) functions from nonlinear potential theory and the viscosity solutions based on generalized pointwise derivatives (jets). Our main result states that in both the elliptic and the parabolic case, the viscosity supersolutions coincide with the potential-theoretic supersolutions.

Key words. p-Laplacian, viscosity solutions, p-superharmonic functions

AMS subject classifications. Primary, 35J70, 35K65; Secondary, 31B99, 35D05

PII. S0036141000372179

1. Introduction.

The objective of this paper is to prove that the viscosity solutions of the p-Laplace equation

$$-\,\mathrm{div}\left(|\nabla u|^{p-2}\nabla u\right) = 0 \tag{1.1}$$

and its parabolic analogue coincide with the usual weak solutions, defined with the aid of test-functions under the integral sign. Our main result is that the viscosity supersolutions are the same as the so-called p-harmonic functions, which are defined through a comparison principle in nonlinear potential theory. In the linear case $p = 2$ the viscosity supersolutions of the Laplace equation $-\Delta u = 0$ are merely the superharmonic functions in classical potential theory. This result and its parabolic counterpart are due to Lions and can be found, for example, in [FIT]. For related results for equations with measurable coefficients see [Je2], and for certain classes of nonlinear equations see [CKSS].

The p-Laplace equation is the Euler–Lagrange equation for the variational integral

$$\int_\Omega |\nabla u(x)|^p\,dx.$$

*Received by the editors May 15, 2000; accepted for publication (in revised form) May 13, 2001; published electronically October 11, 2001. This paper was written while the authors were participating in the program "Potential Theory and Nonlinear Partial Differential Equations" at the Mittag-Leffler Institute in Stockholm.

http://www.siam.org/journals/sima/33-3/37217.html

[†]Department of Mathematics, University of Jyväskylä, P.O.Box 35 (MaD), FIN-40351 Jyväskylä, Finland (peanju@math.jyu.fi). This author's work was supported by the Academy of Finland, project 41964.

[‡]Department of Mathematics, Norwegian University of Science and Technology, N-7491 Trondheim, Norway (lqvist@math.ntnu.no).

[§]Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260-0001 (manfredi@pitt.edu). This author's work was supported in part by NSF grant DMS-9970687 and by the Mittag-Leffler Institute.

Here $1 < p < \infty$ is a fixed exponent, the function u is scalar valued, and Ω is a domain in the n-dimensional Euclidean space. Notice that in the case $p = 2$ (the linear case) we have the Dirichlet integral and the Laplace equation

$$-\Delta u = 0.$$

The p-harmonic equation is the prototype of a class of quasi-linear equations in the form

$$-\operatorname{div} \mathcal{A}_p(x, \nabla u(x)) = 0,$$

and it is fundamental in the nonlinear potential theory; cf. [HKM]. The p-harmonic operator $\operatorname{div}(|\nabla u|^{p-2}\nabla u)$ also appears in many contexts in physics: non-Newtonian fluids (dilatant fluids have $p > 2$, and pseudoplastics have $1 < p < 2$), reaction-diffusion problems, nonlinear elasticity (torsional creep), glaceology ($p = 4/3$), and the thermal radiation of a hydrogen bomb (see [B]), just to mention a few applications.

We will study the mere notion of solutions, subsolutions, and supersolutions. There is something new to be said about this much-investigated basic topic in connection with the so-called viscosity solutions, a modern concept originating in the theory of Hamilton–Jacobi equations. The viscosity solutions have turned out to be indispensable in the case $p = \infty$ (not treated here; see [Je], [JLM]) and quite useful for p finite when it comes to the pointwise interpretation of expedient identities involving second derivatives; cf. [LMS]. We begin with a brief discussion about different definitions of solution.

The solutions with continuous second derivatives (classical solutions) have the advantage of being easy to define: the equation

$$|\nabla u|^2 \Delta u + (p-2) \sum_{i,j=1}^{n} \frac{\partial u}{\partial x_i} \frac{\partial u}{\partial x_j} \frac{\partial^2 u}{\partial x_i \partial x_j} = 0$$

has to hold at each point in the domain Ω. However, when $p \neq 2$ this class of solutions is too restricted for the solvability of the Dirichlet problem. Roughly speaking, this is due to the fact that solutions may not, in general, be of class C^2 at points where $\nabla u = 0$.

To assure the solvability of the Dirichlet boundary value problem, the *p-harmonic functions* are usually defined as (continuous) weak solutions in the Sobolev space $W^{1,p}$. This is the familiar situation with test-functions under the integral sign; see Definition 2.1 below. The uniqueness comes almost for free here. In other words, one has both existence and uniqueness for the Dirichlet boundary value problem.

The definition of *viscosity solutions* is based on pointwise evaluation of the p-harmonic operator

$$\Delta_p u = \operatorname{div}\left(|\nabla u|^{p-2}\nabla u\right),$$

though only for smooth test-functions φ; see section 2 below. The underlying phenomenon is that the classical (of class C^2) sub- and supersolutions are enough to determine the p-harmonic functions, although the latter often are less smooth. This is the content of Corollary 2.6, which states that the viscosity solutions[1] are the p-harmonic functions. To the best of our knowledge, this is a new result.

[1] To indicate the dependence on the exponent p, they are called viscosity p-solutions below.

We have also included a section on the so-called p-parabolic equation

$$u_t - \operatorname{div}(|\nabla u|^{p-2}\nabla u) = 0.$$

According to Corollary 4.5, its viscosity solutions[2] are the p-parabolic functions (continuous weak solutions in a parabolic Sobolev space). The interpretation of this result requires some caution in the range $1 < p < \frac{2n}{n+2}$, because discontinuous "solutions" have to be ruled out. The proof of the equivalence of the parabolic definitions is simpler than in the elliptic situation. We end the paper with a brief discussion of an alternative definition, due to Ishii and Souganidis [IS], for parabolic viscosity solutions in the singular case $1 < p < 2$.

Finally, let us briefly indicate an application. It is to be expected that, at least under suitable conditions, the limit

$$\lim_{t\to\infty} u(x,t)$$

of a p-parabolic function is p-harmonic. Such a theorem has been proved in [Ju] with the viscosity technique, which is advantageous for convergence problems. Our Corollary 2.6 complements the result, making it possible to conclude that the viscosity p-solution, obtained as limit function, is p-harmonic.

2. Definitions. Let Ω denote a domain in $\mathbb{R}^n$. The Sobolev space $W^{1,p}(\Omega)$ consists of all functions $u\,\Omega \to [-\infty,\infty]$ that together with their distributional first derivatives

$$\nabla u = \left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \dots, \frac{\partial u}{\partial x_n}\right)$$

are p-summable. The corresponding local space is denoted by $W^{1,p}_{\mathrm{loc}}(\Omega)$.

Definition 2.1. *We say that a continuous function $u \in W^{1,p}_{\mathrm{loc}}(\Omega)$ is p-harmonic in Ω if*

$$\int_\Omega |\nabla u|^{p-2}\langle \nabla u, \nabla\varphi\rangle\, dx = 0$$

for every $\varphi \in C_0^\infty(\Omega)$. Here $1 < p < \infty$.

By elliptic regularity theory the continuity is redundant in the definition. According to a theorem of Ural'tseva, in the case $p > 2$, later extended by DiBenedetto and Lewis to all $p > 1$, $u \in C^{1,\alpha}_{\mathrm{loc}}(\Omega)$; cf. [Ur], [DB], [Le].

Next we will define the p-superharmonic functions via a comparison principle as in [Li]. Notice immediately that the "fundamental solution"

$$V(x) = |x|^{\frac{p-n}{p-1}}$$

if $1 < p < n$ and

$$V(x) = \log\left(\frac{1}{|x|}\right)$$

[2] To indicate the dependence on the exponent p, they are called parabolic viscosity p-solutions later.

if $p = n$ is, indeed, p-superharmonic in $\mathbb{R}^n$ according to the definition below, although

$$\int_{|x|<1} |\nabla V(x)|^p \, dx = \infty.$$

Any reasonable definition has to include this example. This is taken into account in the following potential-theoretic definition.

Definition 2.2. *The function* $u\,\Omega \to (-\infty, \infty]$ *is called p-superharmonic if*

(i) *u is lower semicontinuous,*

(ii) $u \not\equiv \infty$*, and*

(iii) *u satisfies the comparison principle on each subdomain* $D \Subset \Omega$*: if* $h \in C(\overline{D})$ *is p-harmonic in D and* $u \geq h$ *on* ∂D*, then* $u \geq h$ *in D.*

We have used the notation $D \Subset \Omega$ to indicate that the closure of the domain D is contained in Ω. For $p = 2$, this is the classical definition of superharmonic functions due to Riesz. Note that there are no requirements for the gradient ∇u in Definition 2.2. However, a locally bounded p-superharmonic function u is actually in $W^{1,p}_{\text{loc}}(\Omega)$ and satisfies the inequality

$$\int_\Omega |\nabla u|^{p-2} \langle \nabla u, \nabla \varphi \rangle \, dx \geq 0 \tag{2.1}$$

for every nonnegative test-function $\varphi \in C_0^\infty(\Omega)$. Moreover, the converse is also true for a lower semicontinuous function u in $W^{1,p}_{\text{loc}}(\Omega)$. We refer the reader to [Li] for this fact and more information about the definition and properties of p-superharmonic functions.

Needless to say, *p-subharmonic* functions are defined in an analogous way, so that u is p-subharmonic if and only if $-u$ is p-superharmonic. A function u is p-harmonic if and only if it is both p-superharmonic and p-subharmonic. For the sake of completeness, we mention the *comparison principle* in nonlinear potential theory; see [HKM], [Li].

Comparison principle for p-subharmonic and p-superharmonic functions. Suppose that Ω is a bounded domain, that u is p-subharmonic, and that v is p-superharmonic in Ω. If

$$\limsup_{x \to z} u(x) \leq \liminf_{x \to z} v(x)$$

for all $z \in \partial\Omega$ and if both sides of the inequality are not simultaneously $+\infty$ or $-\infty$, then $u \leq v$ in Ω.

Let us now turn our attention to viscosity solutions. The notion of viscosity solutions requires that the expression

$$\begin{aligned} \Delta_p \varphi &= \operatorname{div}\left(|\nabla \varphi|^{p-2} \nabla \varphi\right) \\ &= |\nabla \varphi|^{p-4} \left[|\nabla \varphi|^2 \Delta \varphi + (p-2) \sum_{i,j=1}^n \frac{\partial \varphi}{\partial x_i} \frac{\partial \varphi}{\partial x_j} \frac{\partial^2 \varphi}{\partial x_i \partial x_j} \right] \end{aligned}$$

be evaluated pointwise for smooth functions φ. This is not a problem when $\nabla \varphi \neq 0$, but the critical points pose additional difficulties, especially in the range $1 < p < 2$. A standard way to deal with singular equations in the theory of viscosity solutions is to use suitable semicontinuous extensions of the operator; cf. [CGG], [CIL]. For the p-Laplacian this approach would allow some false "solutions." For example, $u \equiv$ *constant* would then solve

$$-\Delta_p u(x) = f(x)$$

in the viscosity sense for any continuous function f.

In the following definition, the pointwise evaluation of $\Delta_p \varphi$ is avoided when $\nabla\varphi = 0$. This precaution has no bearing if $p \geq 2$. (Observe that the difficulty with critical points cannot be just defined away, and, in connection with the approximating equation $-\Delta_p v = \varepsilon$, isolated critical points have to be dealt with; see Lemma 3.2 below.) To further motivate the definition, we remark that a function $\varphi \in C^2(\Omega)$ that satisfies $-\Delta_p \varphi(x) = 0$ when $\nabla\varphi(x) \neq 0$ (nothing being said about the possible critical points) is p-harmonic in Ω. This new result is an immediate consequence of Corollary 2.6 below.

DEFINITION 2.3. *The function $u\,\Omega \to (-\infty, \infty]$ is called a viscosity p-supersolution if*

(i) *u is lower semicontinuous,*

(ii) *$u \not\equiv \infty$, and*

(iv) *whenever $x_0 \in \Omega$ and $\varphi \in C^2(\Omega)$ are such that $u(x_0) = \varphi(x_0)$, $u(x) > \varphi(x)$ for $x \neq x_0$, and $\nabla\varphi(x_0) \neq 0$, we have*

$$-\Delta_p \varphi(x_0) \geq 0.$$

Each point x_0 requires its own family of test-functions touching from below, which may very well be empty. It is not difficult to see that condition (iv) can be replaced by the following condition.

(v) The following comparison holds for each subdomain $D \Subset \Omega$: let $\varphi \in C^2(\Omega)$ be such that $\nabla\varphi(x) \neq 0$ and $-\Delta_p \varphi(x) < 0$ in D. If $u \geq \varphi$ on ∂D, then $u \geq \varphi$ in D.

In other words, the comparison is with respect to "smooth strict subsolutions" in Definition 2.3 and with respect to p-harmonic functions in Definition 2.2. Our main result, Theorem 2.5 below, guarantees that both definitions yield the same class of u's. We have come to a fundamental issue about the difference between conditions (iii) and (v). At first sight, condition (v) looks like (iii) in Definition 2.2, especially if one replaces the strict inequality $-\Delta_p \varphi(x) < 0$ by $-\Delta_p \varphi(x) \leq 0$, which is possible a posteriori due to our results. The point is that the comparison in (iii) is with respect to p-harmonic functions that are not necessarily of class C^2, the regularity being merely $C^{1,\alpha}$, while (v) is restricted to C^2-functions. It is in doubt whether one can further restrict the comparison in (iii) to p-harmonic functions h having continuous second derivatives.

An upper semicontinuous function u is a viscosity p-subsolution if $-u$ is a viscosity p-supersolution. A viscosity solution of the equation $-\Delta_p u = 0$ is both a viscosity p-supersolution and p-subsolution.

Remark 2.4. If $p \geq 2$, then $-\Delta_p \varphi(x)$ is well defined also at the critical points of φ, and there is no need to require in (iv) that the gradient of a test-function does not vanish at the point of touching. However, since it turns out that both versions of the definition give the same class of solutions, we have decided to use the one that works also in the singular case $1 < p < 2$.

THEOREM 2.5. *Let $1 < p < \infty$. In a given domain the p-superharmonic functions and the viscosity p-supersolutions are the same.*

COROLLARY 2.6. *Let $1 < p < \infty$. A function is p-harmonic if and only if it is a viscosity p-solution.*

The proof of Theorem 2.5 has two parts. First, we must prove that p-superharmonic functions are viscosity p-supersolutions. This is a rather immediate consequence of the classical comparison principle for p-superharmonic and p-subharmonic functions. Second, we must show that viscosity p-supersolutions are p-superharmonic, that is, they satisfy the comparison principle with respect to p-harmonic functions. This is more delicate since the points at which the gradient vanishes present difficulties in the degenerate as well as in the singular case. Since this comparison principle may be of independent interest, we have stated it by itself. The proof is presented in section 3 below.

THEOREM 2.7 (the comparison principle). *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain, and assume that u is a viscosity p-subsolution and v is a viscosity p-supersolution in Ω. If*

$$\limsup_{x\to z} u(x) \le \liminf_{x\to z} v(x) \tag{2.2}$$

for all $z \in \partial\Omega$ and if both sides of (2.2) *are not simultaneously ∞ or $-\infty$, then $u \le v$ in Ω.*

Proof of Theorem 2.5. Let us first assume that u is p-superharmonic. To show that u is a viscosity p-supersolution, we argue by contradiction and assume that there exists $x_0 \in \Omega$ and $\varphi \in C^2(\Omega)$ such that $u(x_0) = \varphi(x_0)$, $u(x) > \varphi(x)$ for all $x \neq x_0$, $\nabla\varphi(x_0) \neq 0$, and

$$-\Delta_p \varphi(x_0) < 0.$$

By continuity, there exists a radius $r > 0$ such that

$$\begin{cases} -\Delta_p \varphi(x) < 0, \\ \quad \nabla\varphi(x) \neq 0 \end{cases}$$

for every $x \in B_r(x_0)$. Let

$$m = \inf_{|x-x_0|=r} (u(x) - \varphi(x)) > 0,$$

and define $\tilde\varphi = \varphi + m$. Then $\tilde\varphi$ is p-subharmonic in the open set $B_r(x_0)$. Since $\tilde\varphi \le u$ on $\partial B_r(x_0)$, we obtain from the comparison principle for p-superharmonic and p-subharmonic functions that $\tilde\varphi \le u$ in $B_r(x_0)$. However,

$$\tilde\varphi(x_0) = \varphi(x_0) + m > u(x_0),$$

which is a contradiction. Therefore, u must be a viscosity p-supersolution.

For the converse implication, it is enough to check that (iii) holds for viscosity p-supersolutions. This, however, follows immediately from Theorem 2.7 after noticing that by the first half of the proof, every p-harmonic function is a viscosity p-solution. □

3. Proof of the comparison principle. Let the functions u and v satisfy the assumptions in Theorem 2.7. We begin with some simplifications of the general situation.

First reduction (approximation by smooth domains). We may assume, without loss of generality, that the bounded domain Ω is smooth, the function $v \in C^{1,\alpha}(\overline{\Omega})$ is p-harmonic, and $u \le v$ on $\partial\Omega$.

To see this, let us first observe that by (2.2) we can find for any $\epsilon > 0$ a smooth domain $D \Subset \Omega$ such that $u < v + \epsilon$ in $\Omega \setminus D$. By semicontinuity there is a function $\varphi \in C^\infty(\Omega)$ such that

$$u < \varphi < v + \epsilon$$

on ∂D. Next, let h be the unique weak solution to the Dirichlet problem

$$\begin{cases} -\Delta_p h = 0 & \text{in } D, \\ h = \varphi & \text{on } \partial D. \end{cases}$$

In other words, $h \in C(\bar{D}) \cap W^{1,p}(D)$ is p-harmonic in D. Since D is regular, h takes its prescribed continuous boundary values φ in the classical sense.

We have

$$u \le h \le v + \epsilon$$

on ∂D. In fact, it is known that $h \in C^{1,\alpha}(\bar{D})$ (see [Lie]), but we prefer to give an argument which avoids this difficult boundary regularity result. The weaker local regularity $h \in C^{1,\alpha}_{\mathrm{loc}}(D)$ will suffice. To this end, we construct a regular domain $D_1 \Subset D$ such that

$$u - \epsilon \le h \le v + 2\epsilon$$

on ∂D_1 and $u < v + 2\epsilon$ in $\Omega \setminus D_1$. Notice that now we have $h \in C^{1,\alpha}(\overline{D}_1)$, because $h \in C^{1,\alpha}_{\mathrm{loc}}(D)$. If we assume the theorem for regular domains, we get

$$u - \epsilon \le h \le v + 2\epsilon$$

in the whole D_1. Therefore, we conclude that $u \le v + 3\epsilon$ in Ω. Since $\epsilon > 0$ was arbitrary, this is the desired situation. Moreover, since the two cases $u - \epsilon \le h$ and $h \le v + 2\epsilon$ are symmetric, it suffices to prove that $u - \epsilon \le h$ in D_1.

Second reduction (approximation by "regularized" equations). It is enough to prove the comparison principle in the case when v is a weak solution of the equation

$$-\Delta_p v = \varepsilon, \qquad \varepsilon > 0. \tag{3.1}$$

More precisely, suppose that v is a weak solution of (3.1) with smooth boundary values ($v - w \in W^{1,p}_0(\Omega)$ for some $w \in C^{1,\alpha}(\overline{\Omega})$) and Ω is a bounded smooth domain. If u is a viscosity p-subsolution in Ω such that $u(x) \le v(x)$ for all $x \in \partial\Omega$, then we have $u(x) \le v(x)$ for $x \in \Omega$.

Indeed, let us assume that the comparison principle holds in the setting described above, and let u and v be as in the first reduction. If v_ε is the unique weak solution of (3.1) with the boundary condition $v_\varepsilon = v$ on $\partial\Omega$, then by the assumed comparison $u \le v_\varepsilon$ in Ω for every $\varepsilon > 0$. On the other hand, by Lemma 3.1 below, $v_\varepsilon \to v$ locally uniformly. This, in turn, means that $u \le v$ in Ω, which is exactly what we want to prove.

LEMMA 3.1. *Let $v \in W^{1,p}(\Omega)$ be p-harmonic in a bounded domain Ω, and let v_ε be the unique weak solution of the Dirichlet problem*

$$\begin{cases} -\Delta_p v_\varepsilon = \varepsilon & \text{in } \Omega, \\ \quad v_\varepsilon = v & \text{on } \partial\Omega. \end{cases}$$

Then $v_\varepsilon \to v$ locally uniformly in Ω as $\varepsilon \to 0$.

Proof. Take $v - v_\varepsilon \in W_0^{1,p}(\Omega)$ as a test-function in (1.1) and (3.1), and subtract the resulting equations. This yields

$$\begin{aligned}\int_\Omega \big\langle |\nabla v|^{p-2}\nabla v - |\nabla v_\varepsilon|^{p-2}\nabla v_\varepsilon, \nabla v - \nabla v_\varepsilon\big\rangle\, dx &= \varepsilon \int_\Omega (v_\varepsilon - v)\, dx \\ &\le \varepsilon |\Omega|^{\frac{p-1}{p}} \left(\int_\Omega |v - v_\varepsilon|^p\, dx\right)^{\frac{1}{p}} \\ &\le K\varepsilon \left(\int_\Omega |\nabla v - \nabla v_\varepsilon|^p\, dx\right)^{\frac{1}{p}},\end{aligned}$$

where we have used the inequalities of Hölder and Sobolev, and $K = K(p, n, \Omega)$ is some constant depending only on p, n, and Ω.

For $p \ge 2$, it follows easily from the elementary vector inequality [DB, Chapter I]

$$|a - b|^p \le 2^{p-1}\langle |a|^{p-2}a - |b|^{p-2}b, a - b\rangle$$

that

$$\int_\Omega |\nabla v - \nabla v_\varepsilon|^p\, dx \le K\varepsilon^{\frac{p}{p-1}}, \tag{3.2}$$

where $K = K(p, n, \Omega)$. The singular case $1 < p < 2$ is slightly more delicate. Start with the vector inequality [DB, Chapter I]

$$\frac{|a - b|^2}{(|a| + |b|)^{2-p}} \le \gamma \langle |a|^{p-2}a - |b|^{p-2}b, a - b\rangle,$$

where γ depends only on p and n and $a, b \in \mathbb{R}^n$. By Hölder's inequality

$$\int_\Omega |\nabla v - \nabla v_\varepsilon|^p\, dx \le \left(\int_\Omega \frac{|\nabla v - \nabla v_\varepsilon|^2}{(|\nabla v| + |\nabla_\varepsilon|)^{2-p}}\, dx\right)^{\frac{p}{2}} \left(\int_\Omega (|\nabla v| + |\nabla v_\varepsilon|)^p\, dx\right)^{\frac{2-p}{2}},$$

and this time

$$\int_\Omega \frac{|\nabla v - \nabla v_\varepsilon|^2}{(|\nabla v| + |\nabla_\varepsilon|)^{2-p}}\, dx \le \gamma K \varepsilon \left(\int_\Omega |\nabla v - \nabla v_\varepsilon|^p\, dx\right)^{\frac{1}{p}}.$$

Therefore, we have the inequality

$$\int_\Omega |\nabla v - \nabla v_\varepsilon|^p\, dx \le K\varepsilon^p \left(\int_\Omega (|\nabla v| + |\nabla v_\varepsilon|)^p\, dx\right)^{2-p}.$$

Since $\int_\Omega |\nabla v_\varepsilon|^p\, dx$, on the other hand, can be estimated in terms of $\int_\Omega |\nabla v|^p\, dx$ independently of ε for ε small, this implies

$$\int_\Omega |\nabla v - \nabla v_\varepsilon|^p\, dx \le C\,\varepsilon^p \left(1 + \int_\Omega |\nabla v|^p\, dx\right)^{2-p}. \tag{3.3}$$

By estimates (3.2), (3.3), it follows that we have a uniform bound for

$$\int_\Omega |\nabla v_\varepsilon|^p\, dx$$

independent of ε. It then follows from the interior regularity estimates for solutions of (3.1) (cf. [Le], [DB2]) that if we fix a compact set $K \subset \Omega$, we have a uniform bound for $\|v_\varepsilon\|_{C^\alpha(K)}$ for any $1 < p < \infty$. By Ascoli–Arzelà's theorem, there exists a subsequence $\varepsilon_i \to 0$ for which $v_{\varepsilon_i} \to w$ uniformly in K. It follows again from (3.2), (3.3) that indeed $w = v$ and that the full sequence $v_\varepsilon \to v$ locally uniformly in Ω as $\varepsilon \to 0$. □

The weak solutions of (3.1) can be seen as strict supersolutions of (1.1), and this property is of great importance in the proof below of the reduced version of Theorem 2.7. A similar type of approximation argument has been used by Jensen [Je] in connection with the ∞-Laplace equation.

We will need the following result on the "viscosity properties" of weak solutions of (3.1).

Lemma 3.2. *Let $v_\varepsilon \in W^{1,p}(\Omega)$ be a continuous weak solution of the equation $-\Delta_p v_\varepsilon = \varepsilon$ in Ω, and let $x_0 \in \Omega$ and $\varphi \in C^2(\Omega)$ be such that $v_\varepsilon - \varphi$ has a strict local minimum at x_0. Then*

$$\limsup_{\substack{x \to x_0 \\ x \neq x_0}} \left(-\Delta_p \varphi(x)\right) \geq \varepsilon,$$

provided that $\nabla\varphi(x_0) \neq 0$ or x_0 is an isolated critical point.

Remark. We have come to a decisive point. In the case $p \geq 2$ the proof yields that $-\Delta_p\varphi(x_0) \geq \varepsilon$ and that $\nabla\varphi(x) \neq 0$ in some neighborhood of x_0. It is the case $1 < p < 2$ that requires caution, because $-\Delta_p\varphi(x)$ is undetermined at the critical points (which may be encountered).

Proof. Suppose that the assertion is not true, that is, there is $r > 0$ such that

$$\nabla\varphi(x) \neq 0 \quad \text{and} \quad -\Delta_p\varphi(x) < \varepsilon,$$

when $0 < |x - x_0| < r$. After a translation, we may assume that $x_0 = 0$. Take any nonnegative test-function $\phi \in C_0^\infty(B_r)$, and integrate over the annulus $\rho < |x| < r$. (The auxiliary $\rho > 0$ can be skipped if $\nabla\varphi(0) \neq 0$.) According to Gauss's theorem,

$$\begin{aligned} -\oint_{|x|=\rho} \phi|\nabla\varphi|^{p-2}\langle\nabla\varphi, \tfrac{x}{\rho}\rangle\, dS &= \int_{\rho<|x|<r} \operatorname{div}(\phi|\nabla\varphi|^{p-2}\nabla\varphi)\, dx \\ &= \int_{\rho<|x|<r} |\nabla\varphi|^{p-2}\langle\nabla\varphi, \nabla\phi\rangle\, dx + \int_{\rho<|x|<r} \phi(\Delta_p\varphi)\, dx. \end{aligned}$$

The flux approaches 0 as $\rho \to 0_+$. Indeed,

$$\left| \oint_{|x|=\rho} \phi|\nabla\varphi|^{p-2}\langle\nabla\varphi, \tfrac{x}{\rho}\rangle\, dS \right| \leq \|\phi\|_\infty \|\nabla\varphi\|_\infty^{p-1}\omega_{n-1}\rho^{n-1},$$

where $\omega_{n-1}\rho^{n-1}$ is the area of the sphere of radius ρ. By the antithesis we have

$$\int_{\rho<|x|<r} \phi(\Delta_p\varphi)\, dx \geq -\varepsilon \int_{\rho<|x|<r} \phi\, dx \geq -\varepsilon \int_{B_r} \phi\, dx.$$

Therefore, we obtain

$$\int_{B_r} |\nabla\varphi|^{p-2}\langle\nabla\varphi, \nabla\phi\rangle\, dx = \lim_{\rho\to 0} \int_{\rho<|x|<r} |\nabla\varphi|^{p-2}\langle\nabla\varphi, \nabla\phi\rangle\, dx \leq \varepsilon \int_{B_r} \phi\, dx.$$

Thus φ is a weak subsolution of (3.1). We finish by using the comparison principle as in the proof of Theorem 2.5 presented in section 2. $\square$

After the reductions made above, it suffices to prove the following version of the comparison principle [see (3.4) below]. In the proof we use the notation

$$F(\eta, X) = -|\eta|^{p-2}\left[\operatorname{trace}(X) + (p-2)\left\langle X\frac{\eta}{|\eta|}, \frac{\eta}{|\eta|}\right\rangle\right] \tag{3.4}$$

when $\eta \neq 0$ is a vector in $\mathbb{R}^n$ and $X \in S_n$, where S_n denotes the class of real symmetric $n \times n$ matrices. For a smooth function φ we clearly have

$$F(\nabla\varphi(x), D^2\varphi(x)) = -\operatorname{div}(|\nabla\varphi(x)|^{p-2}\nabla\varphi(x))$$

when $\nabla\varphi(x) \neq 0$. Here $D^2\varphi = (\frac{\partial^2\varphi}{\partial x_i \partial x_j})_{n\times n}$ is the Hessian matrix of φ.

PROPOSITION 3.3. *Suppose that $\Omega \subset \mathbb{R}^n$ is a smoothly bounded domain, u is a viscosity p-subsolution, and $v \in C^{1,\alpha}(\overline{\Omega})$ is a weak solution of $-\Delta_p v = \varepsilon$ in Ω such that $u \leq v$ on $\partial\Omega$. Then $u \leq v$ in Ω.*

Proof. Without loss of generality, we may assume that $\varepsilon = 1$. We argue by contradiction and assume that $u - v$ has an interior maximum, that is,

$$\sup_{\Omega}(u-v) > \sup_{\partial\Omega}(u-v). \tag{3.5}$$

Consider the functions

$$w_j(x,y) = u(x) - v(y) - \Psi_j(x,y), \quad j = 1, 2, \ldots,$$

where

$$\Psi_j(x,y) = \tfrac{j}{q}|x-y|^q, \qquad q > \max\left\{\frac{p}{p-1}, 2\right\},$$

and let (x_j, y_j) be a maximum of w_j relative to $\overline{\Omega} \times \overline{\Omega}$. By (3.5) and Proposition 3.7 in [CIL], we see that for j sufficiently large, (x_j, y_j) is an interior point. Since

$$u(x) - v(y) - \Psi_j(x,y) \leq u(x_j) - v(y_j) - \Psi_j(x_j, y_j)$$

for all $x, y \in \Omega$, we obtain by choosing $x = x_j$ that

$$v(y) \geq -\Psi_j(x_j, y) + v(y_j) + \Psi_j(x_j, y_j)$$

for all $y \in \Omega$. Let us denote

$$\phi_j(y) = -\Psi_j(x_j, y) + v(y_j) + \Psi_j(x_j, y_j) - \frac{1}{q+1}|y - y_j|^{q+1}.$$

Then, clearly, $v - \phi_j$ has a strict local minimum at y_j, and thus

$$\limsup_{\substack{y \to y_j \\ y \neq y_j}} (-\Delta_p \phi_j(y)) \geq 1$$

by Lemma 3.2. This implies that $x_j \neq y_j$. Indeed, if $x_j = y_j$, then a direct computation shows that $-\Delta_p \phi_j(y) \to 0$ as $y \to y_j$, which is a contradiction.

The rest of the proof is now a rather standard application of the maximum principle for semicontinuous functions (also known as the theorem on sums) in [CIL]. Since (x_j, y_j) is a local maximum point of $w_j(x, y)$, we conclude that there exist symmetric $n \times n$ matrices X_j, Y_j such that

$$\begin{aligned}(D_x\Psi_j(x_j, y_j), X_j) &\in \overline{J}^{2,+}u(x_j),\\ (-D_y\Psi_j(x_j, y_j), Y_j) &\in \overline{J}^{2,-}v(y_j),\end{aligned}$$

and

$$\begin{pmatrix} X_j & 0 \\ 0 & -Y_j \end{pmatrix} \le D^2\Psi_j(x_j, y_j) + \frac{1}{j}\left[D^2\Psi_j(x_j, y_j)\right]^2. \tag{3.6}$$

Here $\overline{J}^{2,+}u(x_j)$ and $\overline{J}^{2,-}v(y_j)$ are the closures of the second order superjet of u at x_j and the second order subjet of v at y_j, respectively. We refer the reader to [C] and [CIL] for the definition and properties of jets.

Observe that since $D^2\Psi_j$ annihilates vectors of the form $\binom{\xi}{\xi}$, we obtain from (3.6) that

$$X_j \le Y_j$$

in the sense of matrices, that is, $\langle (Y_j - X_j)\xi, \xi\rangle \ge 0$ for all $\xi \in \mathbb{R}^n$.

Let us now finish the proof. It is well known (see [CIL]) that for equations that are continuous in each variable, viscosity solutions can be defined using jets instead of test-functions as in Definition 2.3. Since $x_j \neq y_j$, we have that

$$\eta_j \equiv D_x\Psi_j(x_j, y_j) = -D_y\Psi_j(x_j, y_j) \neq 0.$$

This means that

$$(\eta, X) \mapsto F(\eta, X),$$

where F is given by (3.4) and is continuous in a neighborhood of the points (η_j, X_j) and (η_j, Y_j), and we may use the equivalent definition involving jets. Since u is a subsolution of (1.1), we obtain that

$$-|\eta_j|^{p-2}\left[\operatorname{trace}(X_j) + (p-2)\left\langle X_j\frac{\eta_j}{|\eta_j|}, \frac{\eta_j}{|\eta_j|}\right\rangle\right] \le 0.$$

On the other hand, since $\eta_j \neq 0$, by the definition of $\overline{J}^{2,-}$ Lemma 3.2 implies that

$$-|\eta_j|^{p-2}\left[\operatorname{trace}(Y_j) + (p-2)\left\langle Y_j\frac{\eta_j}{|\eta_j|}, \frac{\eta_j}{|\eta_j|}\right\rangle\right] \ge 1.$$

Hence

$$\begin{aligned}0 < 1 \le &- |\eta_j|^{p-2}\left[\operatorname{trace}(Y_j) + (p-2)\left\langle Y_j\frac{\eta_j}{|\eta_j|}, \frac{\eta_j}{|\eta_j|}\right\rangle\right]\\ &+ |\eta_j|^{p-2}\left[\operatorname{trace}(X_j) + (p-2)\left\langle X_j\frac{\eta_j}{|\eta_j|}, \frac{\eta_j}{|\eta_j|}\right\rangle\right]\\ \le&\ 0,\end{aligned}$$

where the last inequality follows from the fact $X_j \leq Y_j$. This contradiction means that our initial assumption (3.5) cannot hold, and, therefore,

$$\sup_{\Omega}(u - v) = \sup_{\partial\Omega}(u - v) \leq 0$$

as claimed. □

4. The parabolic case. The p-parabolic equation

$$u_t - \operatorname{div}\left(|\nabla u|^{p-2}\nabla u\right) = 0, \tag{4.1}$$

where $u = u(x,t)$, has the p-harmonic equation as its stationary equation. Let us introduce some notation. Let

$$Q = (a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_n, b_n)$$

denote a parallelepiped, and consider the "space-time box"

$$Q_{t_1,t_2} = Q \times (t_1, t_2)$$

in the (x,t)-space. Its parabolic boundary is

$$\partial_{par} Q = \left(\overline{Q} \times \{t_1\}\right) \cup \left(\partial Q \times (t_1, t_2]\right).$$

It consists of the bottom and the lateral sides, but the interior points of the top are excluded.

In order to describe the appropriate function space, we introduce the abbreviation

$$V^p(t_1, t_2; Q) = C\left(t_1, t_2; L^2(Q)\right) \cap L^p\left(t_1, t_2; W^{1,p}(Q)\right).$$

Thus $u \in V^p(t_1, t_2; Q)$ implies that the mapping

$$t \mapsto \int_Q |u(x,t)|^2 \, dx$$

is continuous in $[t_1, t_2]$, the Sobolev derivative

$$\nabla u(x,t) = \left(\frac{\partial u(x,t)}{\partial x_1}, \ldots, \frac{\partial u(x,t)}{\partial x_n}\right)$$

exists for almost every t in $[t_1, t_2]$, and the integral

$$\int_{t_1}^{t_2} \int_Q \left(u^2 + |\nabla u|^p\right) \, dt \, dx$$

is finite.

Definition 4.1. *Let $\mathcal{O}$ be a domain in $\mathbb{R}^n \times \mathbb{R}$, and suppose that the function $u\, \mathcal{O} \to \mathbb{R}$ is continuous and belongs to $V^p(t_1, t_2; Q)$ whenever the closure of Q_{t_1,t_2} is comprised in $\mathcal{O}$. We say that u is p-parabolic in $\mathcal{O}$ if*

$$\int\int_{\mathcal{O}} (-u\varphi_t + |\nabla u|^{p-2}\langle \nabla u, \nabla \varphi \rangle) \, dt \, dx = 0 \tag{4.2}$$

for all test-functions $\varphi \in C_0^\infty(\mathcal{O})$.

By parabolic regularity theory, the continuity is a redundant requirement in the definition if $p > \frac{2n}{n+2}$. The case $1 < p \leq \frac{2n}{n+2}$ is not fully understood. It is known that there exist locally unbounded u's satisfying (4.1) in a weaker sense than described above. We refer to [DB, Chapter XII]. Such weak solutions cannot be viscosity solutions, since the latter are both upper and lower semicontinuous by definition and hence locally bounded. Thus our characterization of the p-parabolic viscosity solutions discards such discontinuous weak solutions.

The definition below includes the celebrated Barenblatt solution, which for $p > 2$ is given as

$$\mathcal{B}_p(x,t) = \begin{cases} t^{-\frac{n}{\lambda}} \left\{ C - \frac{p-2}{2} \lambda^{\frac{1}{1-p}} \left(\frac{|x|}{t^{1/\lambda}} \right)^{\frac{p}{p-1}} \right\}_+^{\frac{p-1}{p-2}}, & t > 0, \\ 0, & t \leq 0, \end{cases}$$

when $(x,t) \neq (0,0)$. Here $\lambda = n(p-2) + p$, and C is a positive constant. With the definition $\mathcal{B}_p(0,0) = \infty$, the Barenblatt solution is p-superparabolic in the whole $\mathbb{R}^n \times \mathbb{R}$, although

$$\int_{-1}^{1} \int_{|x|<1} |\nabla \mathcal{B}_p(x,t)|^p \, dx \, dt = \infty.$$

Definition 4.2. *A function $u\ : \mathcal{O} \to \mathbb{R} \cup \{\infty\}$ is p-superparabolic if*

(i) *u is lower semicontinuous,*

(ii′) *u is finite in a dense subset of $\mathcal{O}$, and*

(iii) *u satisfies the comparison principle on each box Q_{t_1,t_2} with closure in $\mathcal{O}$: if $h \in C(\overline{Q}_{t_1,t_2})$ is p-parabolic in Q_{t_1,t_2} such that $h \leq u$ on the parabolic boundary of Q_{t_1,t_2}, then $h \leq u$ in Q_{t_1,t_2}.*

We refer to [KL] for a detailed discussion on the properties of the p-superparabolic functions.

Let us next turn to the definition of viscosity solutions of (4.1). Due to the presence of the time derivative u_t in the equation, we cannot exclude test-functions with vanishing spatial gradient $\nabla\varphi(x,t)$ at the point of touching like we did with the p-harmonic equation. As in the elliptic case, the equation is singular only in the range $1 < p < 2$, but we have chosen again not to distinguish between the two cases.

Definition 4.3. *A function $u\ : \mathcal{O} \to \mathbb{R} \cup \{\infty\}$ is a parabolic viscosity p-supersolution if*

(i) *u is lower semicontinuous,*

(ii′) *u is finite in a dense subset of $\mathcal{O}$, and*

(iv) *whenever $(x_0,t_0) \in \mathcal{O}$ and $\varphi \in C^2(\mathcal{O})$ are such that $u(x_0,t_0) = \varphi(x_0,t_0)$, $u(x,t) > \varphi(x,t)$ for $(x,t) \in \mathcal{O} \cap \{t < t_0\}$, and $\nabla\varphi(x,t) \neq 0$ if $x \neq x_0$, we have*

$$\limsup_{\substack{(x,t)\to(x_0,t_0) \\ t<t_0, x\neq x_0}} \Big(\varphi_t(x,t) - \Delta_p \varphi(x,t) \Big) \geq 0.$$

The concepts of *parabolic viscosity p-subsolution* and *parabolic viscosity p-solution* are defined analogously. Notice that the parabolic viscosity p-solutions are continuous by definition. As in the elliptic case, the precaution about $\nabla\varphi \neq 0$ can be ignored for $p \geq 2$.

What is to happen in the future will have no influence on the present time. This phenomenon, typical of parabolic equations, was taken into account in the definition above: the test-function is forced to be under the function u only up to the time t_0 of testing. This yields the same concept as a definition with no emphasis on the special role of the time variable; cf. [Ju].

THEOREM 4.4. *Let $1 < p < \infty$. In a given domain, the p-superparabolic functions and the parabolic viscosity p-supersolutions are the same.*

COROLLARY 4.5. *Let $1 < p < \infty$. A continuous function is a parabolic viscosity p-solution if and only if it is p-parabolic.*

The proof of Theorem 4.4 is virtually the same as its elliptic counterpart. To show that a p-superparabolic function is a parabolic viscosity p-supersolution, one needs to consider space-time boxes instead of balls and use the comparison principle for p-superparabolic and p-subparabolic functions; see [KL]. For the converse, the comparison principle for viscosity solutions is needed (Theorem 4.10 below).

LEMMA 4.6. *Every p-superparabolic function u is a parabolic viscosity p-supersolution.*

Proof. We argue by contradiction and assume that there exist $\varphi \in C^2(\mathcal{O})$ and $r > 0$ such that $u(0,0) = \varphi(0,0)$, $u(x,t) > \varphi(x,t)$ for all $(x,t) \in \mathcal{O} \cap \{t < 0\}$, $\nabla\varphi(x,t) \neq 0$ when $x \neq 0$, and

$$\varphi_t(x,t) - \Delta_p \varphi(x,t) < 0 \tag{4.3}$$

whenever $(x,t) \in Q_r \cup \{x \neq 0\}$, where $Q_r \equiv B_r(0) \times (-r, 0)$. Then for every nonnegative $\phi \in C_0^\infty(Q_r)$, we obtain using (4.3) and Gauss's theorem as in the proof of Lemma 3.4 that

$$\begin{aligned}
&- \iint_{Q_r} |\nabla\varphi|^{p-2} \langle \nabla\varphi, \nabla\phi \rangle \, dx\, dt \\
&= \lim_{\rho \to 0} \left[\iint_{Q_r \setminus \{|x| \leq \rho\}} \phi \left(\Delta_p \varphi \right) \, dx\, dt - \iint_{Q_r \setminus \{|x| \leq \rho\}} \operatorname{div}(\phi |\nabla\varphi|^{p-2} \nabla\varphi) \, dx\, dt \right] \\
&\geq \lim_{\rho \to 0} \left[\iint_{Q_r \setminus \{|x| \leq \rho\}} \phi \varphi_t \, dx\, dt + \int_{-r}^{0} \oint_{|x| = \rho} \phi |\nabla\varphi|^{p-2} \langle \nabla\varphi, \tfrac{x}{\rho} \rangle \, dS\, dt \right] \\
&= - \iint_{Q_r} \phi_t \varphi \, dx\, dt.
\end{aligned}$$

This implies that φ is p-subparabolic in Q_r; see [KL]. To conclude, we proceed as in the elliptic case, and apply the comparison principle for p-superparabolic and p-subparabolic functions from [KL] to the functions u and $\varphi + m$, where

$$m = \inf_{\partial_{par} Q_r} (u - \varphi) > 0.$$

This gives the desired contradiction. □

Due to the fact that the time derivative u_t appears as a linear term in (4.1), Theorem 4.7 below is easier to prove than the elliptic comparison principle, Theorem 2.7. In fact, the nonsingular case $p \geq 2$ follows from a very general result in [C]. Since the basic idea of the proof is quite similar to the one of the elliptic case, we will be somewhat sketchy.

THEOREM 4.7. *Let $\Omega_T = \Omega \times (0, T)$, where $\Omega \subset \mathbb{R}^n$ is a bounded domain, and assume that u is a parabolic viscosity p-subsolution and v is a parabolic viscosity p-supersolution in Ω_T. If $u \leq v$ on the parabolic boundary of Ω_T, then $u \leq v$ in Ω_T.*

Proof. For simplicity, we assume that u is bounded from above and v is bounded from below in $\overline{\Omega} \times [0,T]$. Since the proof is by contradiction, we assume that

$$\sup_{\Omega_T}(u-v) > \sup_{\partial_{par}\Omega_T}(u-v), \tag{4.4}$$

where $\partial_{par}\Omega_T$ denotes the parabolic boundary of Ω_T. By using the standard trick of replacing v by $v(x,t) + \frac{\varepsilon}{T-t}$ for small $\varepsilon > 0$, we may assume that v is a strict supersolution of (4.1) and $v(x,t) \to \infty$ as $t \to T$.

Let (x_j, y_j, t_j, s_j) be a maximum point of

$$w_j(x,y,t,s) = u(x,t) - v(y,s) - \Psi_j(x,y,t,s)$$

relative to $\overline{\Omega} \times \overline{\Omega} \times [0,T]$. Here

$$\Psi_j(x,y,t,s) = \frac{j}{q}|x-y|^q + \frac{j}{2}(t-s)^2, \quad q > \max\left\{\frac{p}{p-1}, 2\right\}.$$

By (4.4) and Proposition 3.7 in [CIL], we have that $(x_j, y_j, t_j, s_j) \in \Omega \times \Omega \times (0,T) \times (0,T)$ for j large enough. We distinguish between two cases.

Case 1. $x_j = y_j$.

By the choice of the point (x_j, y_j, t_j, s_j) we have

$$v(y,s) \geq -\Psi_j(x_j, y, t_j, s) + \Psi_j(x_j, y_j, t_j, s_j) + v(y_j, s_j)$$

for all $(y,s) \in \Omega \times [0,T]$; that is,

$$\phi(y,s) \equiv -\Psi_j(x_j, y, t_j, s) + \Psi_j(x_j, y_j, t_j, s_j) + v(y_j, s_j) - \frac{1}{q+1}|y-y_j|^{q+1}$$

is touching v from below at (y_j, s_j). Since v is a strict supersolution of (4.1) and $x_j = y_j$, we obtain after straightforward computations

$$0 < \frac{\varepsilon}{(T-s_j)^2} \leq \limsup_{\substack{(y,s)\to(y_j,s_j)\\ s<s_j, y\neq y_j}} \Big(\phi_s(y,s) - \Delta_p \phi(y,s)\Big) = j(t_j - s_j). \tag{4.5}$$

Similarly, we see that

$$\theta(x,t) \equiv \Psi_j(x, y_j, t, s_j) - \Psi_j(x_j, y_j, t_j, s_j) + u(x_j, t_j) + \frac{1}{q+1}|x-x_j|^{q+1}$$

is a good test-function for u at the point (x_j, t_j), and hence

$$0 \geq \liminf_{\substack{(x,t)\to(x_j,t_j)\\ t<t_j, x\neq x_j}} \Big(\theta_t(x,t) - \Delta_p \theta(x,t)\Big) = j(t_j - s_j). \tag{4.6}$$

Subtracting (4.6) from (4.5) gives

$$0 < \frac{\varepsilon}{(T-s_j)^2} \leq j(t_j - s_j) - j(t_j - s_j) = 0,$$

which is a contradiction.

Case 2. $x_j \neq y_j$.

As in the elliptic case, we may use the definition with jets. Using the maximum principle for semicontinuous functions and Lemma 3.5 from [OS], we infer that there exist $X_j, Y_j \in S_n$ such that

$$\begin{aligned}(D_t\Psi_j, D_x\Psi_j, X_j) &\in \overline{\mathcal{P}}^{2,+}u(x_j,t_j),\\ (-D_s\Psi_j, -D_y\Psi_j, Y_j) &\in \overline{\mathcal{P}}^{2,-}v(y_j,s_j),\end{aligned}$$

and

$$X_j \leq Y_j. \tag{4.7}$$

Here all the derivatives of Ψ_j are evaluated at the point (x_j, y_j, t_j, s_j). For the definition and properties of the parabolic jets $\mathcal{P}^{2,+}u$ and $\mathcal{P}^{2,-}v$ and their closures, we refer the reader to [C] and [CIL].

Let us now finish the proof. Since u is a subsolution and v is a strict supersolution, we obtain

$$0 < \frac{\varepsilon}{(T-s_j)^2} \leq -D_s\Psi_j + F(-D_y\Psi_j, Y_j) - D_t\Psi_j - F(D_x\Psi_j, X_j) \leq 0,$$

which is a contradiction. Here the last inequality follows from (4.7) after we notice that

$$\begin{aligned}D_t\Psi_j(x,y,t,s) &= -D_s\Psi_j(x,y,t,s),\\ D_x\Psi_j(x,y,t,s) &= -D_y\Psi_j(x,y,t,s),\end{aligned}$$

by the choice of Ψ_j. This shows that (4.4) cannot hold, and we are done. □

Remark 4.8. In [CGG], Chen, Giga, and Goto obtained a general comparison theorem for mean curvature flow-type equations. Those equations are singular at the points where the spatial gradient vanishes, but the nature of the singularity is different from that of the p-parabolic equation. Roughly speaking, the mean curvature flow equation has a bounded discontinuity at the points of singularity, whereas the p-parabolic equation behaves like $O(|\nabla u|^{p-2})$ near those points.

We finish the paper with a brief discussion on another possible definition for viscosity solutions of (4.1) in the singular case $1 < p < 2$. This approach is due to Ishii and Souganidis [IS], and in connection with the p-parabolic equation it has been used by Ohnuma and Sato; cf. [OS].

Let us introduce some notation. We set

$$\mathcal{F} = \Big\{ f \in C^2([0,\infty)) \;\; f(0) = f'(0) = f''(0) = 0,\; f''(r) > 0 \text{ for all } r > 0, \text{ and } \lim_{x\to 0_+}(-\Delta_p f(|x|)) = 0 \Big\}$$

and

$$\Sigma = \Big\{ \sigma \in C^1(\mathbb{R}) \;\; \sigma \text{ is even}, \sigma(0) = \sigma'(0) = 0, \text{ and } \sigma(r) > 0 \text{ for all } r \neq 0 \Big\}.$$

DEFINITION 4.9. *A function $\varphi \in C^2(\mathcal{O})$ is admissible if for any $(\hat{x}, \hat{t}) \in \mathcal{O}$ with $\nabla\varphi(\hat{x}, \hat{t}) = 0$ there are $\delta > 0$, $f \in \mathcal{F}$, and $\sigma \in \Sigma$ such that*

$$|\varphi(x,t) - \varphi(\hat{x},\hat{t}) - \varphi_t(\hat{x},\hat{t})(t-\hat{t})| \leq f(|x-\hat{x}|) + \sigma(t-\hat{t})$$

for all $(x,t) \in B_\delta(\hat{x}) \times (\hat{t}-\delta, \hat{t}+\delta)$.

Notice that if $\nabla\varphi \neq 0$, then φ is automatically admissible. The idea of introducing the admissible class is, roughly speaking, to have a good a priori control on the behavior of a test-function at the singular points. For lack of a better terminology, we call the solutions defined using the class of admissible test-functions relaxed viscosity solutions.

DEFINITION 4.10. *A function $u\ : \mathcal{O} \to \mathbb{R} \cup \{\infty\}$ is a relaxed viscosity p-supersolution if*

(i) *u is lower semicontinuous,*

(ii′) *u is finite in a dense subset of $\mathcal{O}$, and*

(v) *for all admissible $\varphi \in C^2(\mathcal{O})$ and all local minimum points (x,t) of $u-\varphi$ in $\mathcal{O}$*

$$\begin{cases} \varphi_t(x,t) - \Delta_p\varphi(x,t) \geq 0 & \textit{if } \nabla\varphi(x,t) \neq 0, \\ \varphi_t(x,t) \geq 0 & \textit{if } \nabla\varphi(x,t) = 0. \end{cases}$$

We have taken the liberty to modify the definition given in [IS], [OS] in order to make a comparison with Definition 4.3 easier. In particular, the original definition of Ishii and Souganidis was formulated without the semicontinuity assumption (i), and thus it does not imply continuity for the solutions.

LEMMA 4.11. *Every parabolic viscosity p-supersolution is a relaxed viscosity p-supersolution.*

Proof. We argue by contradiction and assume that there exist an admissible test-function $\varphi \in C^2(\mathcal{O})$ and $(x_0,t_0) \in \mathcal{O}$ such that $u-\varphi$ has a local minimum at (x_0,t_0), $\nabla\varphi(x_0,t_0) = 0$, and

$$\varphi_t(x_0,t_0) < 0. \tag{4.8}$$

Let $f \in \mathcal{F}$, $\sigma \in \Sigma$, and $\delta > 0$ be such that

$$|\varphi(x,t) - \varphi(x_0,t_0) - \varphi_t(x_0,t_0)(t-t_0)| \leq f(|x-x_0|) + \sigma(t-t_0) \tag{4.9}$$

for all $(x,t) \in B_\delta(x_0) \times (t_0-\delta, t_0+\delta)$. Following the ideas in [IS], we approximate σ by a sequence $\sigma_k \in C^2(\mathbb{R})$ satisfying

$$\begin{cases} \sigma_k(0) = \sigma_k'(0) = 0 \quad \text{for each } k = 1, 2, \ldots, \\ \sigma_k(r) \to \sigma(r), \ \sigma_k'(r) \to \sigma'(r) \quad \text{locally uniformly,} \end{cases}$$

and we denote

$$\begin{aligned} \phi(x,t) &= u(x_0,t_0) + \varphi_t(x_0,t_0)(t-t_0) - 2f(|x-x_0|) - 2\sigma(t-t_0), \\ \phi_k(x,t) &= u(x_0,t_0) + \varphi_t(x_0,t_0)(t-t_0) - 2f(|x-x_0|) - 2\sigma_k(t-t_0). \end{aligned}$$

Observe that (4.9) implies that $u-\phi$ has a strict local minimum at (x_0,t_0). Since $\sigma_k \to \sigma$ locally uniformly, we can find a sequence $(x_k,t_k) \to (x_0,t_0)$ such that $u-\phi_k$ has a local minimum at (x_k,t_k). Moreover, by modifying ϕ_k if necessary, we may assume that this local minimum is, in fact, strict. Hence ϕ_k can be used as a test-function in Definition 4.3, and we obtain

$$\limsup_{\substack{(x,t)\to(x_k,t_k) \\ t<t_k,\, x\neq x_0}} \Big((\phi_k)_t(x,t) - \Delta_p\phi_k(x,t) \Big) \geq 0$$

for each $k \in \mathbb{N}$. However, a direct computation yields

$$(\phi_k)_t(x,t) - \Delta_p \phi_k(x,t) = \varphi_t(x_0,t_0) - 2\sigma_k'(t-t_0) - 2^{p-1}\Delta_p f(|x-x_0|) < 0$$

if (x,t) is sufficiently close to (x_0,t_0) and $x \neq x_0$. Here we used (4.8), the definition of $\mathcal{F}$, and the fact that $\sigma_k'(0) = 0$. This contradiction shows that the antithesis was wrong, and the lemma is now proved. □

In [OS], Ohnuma and Sato proved a comparison principle for the relaxed viscosity p-supersolutions and subsolutions. In the light of Lemmas 4.6 and 4.11, this means that relaxed viscosity p-supersolutions satisfy (p-iii) in Definition 4.2, and hence they are precisely the parabolic viscosity p-supersolutions.

Acknowledgments. We are grateful to Eero Saksman for pointing out an error in an earlier version of the manuscript. We thank the anonymous referee for the careful reading of the manuscript and useful suggestions and comments.

REFERENCES

[B] G. I. Barenblatt, *On selfsimilar motions of compressible fluids in a porous medium*, Prikl. Mat. Mekh., 16 (1952), pp. 679–698 (in Russian).

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

[C] M. G. Crandall, *Viscosity solutions: A primer*, in Viscosity Solutions and Applications (Montecatini Terme, 1995), Lecture Notes in Math. 1660, Springer-Verlag, Berlin, 1997, pp. 1–43.

[CIL] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[CKSS] M. Crandall, M. Kocan, P. Soravia, and A. Swiech, *On the equivalence of various weak notions of solutions of elliptic PDEs with measurable ingredients*, in Progress in Elliptic and Parabolic Partial Differential Equations, Pitman Res. Notes Math. Ser. 350, A. Alvino and P. Buoncore, eds., Longman, Harlow, UK, 1996, pp. 136–162.

[DB] E. DiBenedetto, *Degenerate Parabolic Equations*, Springer-Verlag, New York, 1993.

[DB2] E. DiBenedetto, *$C^{1+\alpha}$-local regularity of weak solutions of degenerate elliptic equations*, Nonlinear Anal., 7 (1983), pp. 827–850.

[FIT] I. Fukuda, H. Ishii, and M. Tsutsumi, *Uniqueness of solutions to the Cauchy problem for $u_t + u\Delta u + \gamma|\nabla u|^2 = 0$*, Differential Integral Equations, 6 (1993), pp. 1231–1252.

[HKM] J. Heinonen, T. Kilpeläinen, and O. Martio, *Nonlinear Potential Theory of Degenerate Elliptic Equations*, Oxford University Press, Oxford, UK, 1993.

[IS] H. Ishii and P. E. Souganidis, *Generalized motion of noncompact hypersurfaces with velocity having arbitrary growth on the curvature tensor*, Tohoku Math. J. (2), 47 (1995), pp. 227–250.

[Je] R. Jensen, *Uniqueness of Lipschitz extensions: Minimizing the sup norm of the gradient*, Arch. Ration. Mech. Anal., 123 (1993), pp. 51–74.

[Je2] R. Jensen, *Uniformly elliptic PDEs with bounded, measurable coefficients*, J. Fourier Anal. Appl., 2 (1996), pp. 237–259.

[Ju] P. Juutinen, *On the definition of viscosity solutions for parabolic equations*, Proc. Amer. Math. Soc., 129 (2001), pp. 2907–2911.

[JLM] P. Juutinen, P. Lindqvist, and J. Manfredi, *The ∞-eigenvalue problem*, Arch. Ration. Mech. Anal., 148 (1999), pp. 89–105.

[KL] 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.

[Le] J. Lewis, *Regularity of the derivatives of solutions to certain elliptic equations*, Indiana Univ. Math. J., 32 (1983), pp. 849–858.

[Li] P. Lindqvist, *On the definition and properties of p-superharmonic functions*, J. Reine Angew. Math., 365 (1986), pp. 67–79.

[Lie] G. M. Lieberman, *Boundary regularity for solutions of degenerate elliptic equations*, Nonlinear Anal., 12 (1988), pp. 1202–1219.

[LMS] P. Lindqvist, J. Manfredi, and E. Saksman, *Superharmonicity of nonlinear ground states*, Rev. Mat. Iberoamericana, 16 (2000), pp. 17–28.

[OS] S. Ohnuma and K. Sato, *Singular degenerate parabolic equations with applications to the p-Laplace diffusion equation*, Comm. Partial Differential Equations, 22 (1997), pp. 381–411.

[Ur] N. Ural'tseva, *Degenerate quasilinear elliptic systems*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. POMI, 7 (1968), pp. 184–222 (in Russian).

SIAM J. Math. Anal.
Vol. 33, No. 3, pp. 718–750

THE LOCAL BEHAVIOR OF THE SOLUTION OF THE RADIOSITY EQUATION AT THE VERTICES OF POLYHEDRAL DOMAINS IN $\mathbb{R}^3$*

OLAF HANSEN†

Abstract. This article studies the regularity of the solution of the radiosity equation on polyhedral surfaces in $\mathbb{R}^3$. We examine the localized equation near the vertices and characterize the smoothness of its solution in terms of weighted Sobolev spaces. We also consider the effect of shadows and prove local $\mathcal{C}^1$-regularity in the case of a piecewise $\mathcal{C}^1$-emissivity function.

Key words. radiosity equation, regularity results, Mellin transformation

AMS subject classifications. 45P05, 44A15, 45Exx, 68U05

PII. S0036141000378103

1. Introduction.

In this paper we study the properties of the solution of the radiosity equation over the boundary $S := \partial\Omega$ of a polyhedron $\Omega \subset \mathbb{R}^3$. This implies that the boundary S is a finite union of triangles.

The solution $\overline{u} : S \to \mathbb{R}$ of the radiosity equation describes the outgoing radiation at every point of the surface of S if the emissivity $E : S \to \mathbb{R}$, i.e., the sources of illumination and their brightness, on S is known. A further assumption is that the walls (S) behave like Lambertian diffuse reflectors. For many surfaces, i.e., paper or walls which are not glossy, this assumption is fulfilled. Even if there are mirrors or glossy surfaces, there exists a number of techniques for including their effects on the overall radiosity in a second computational step after the calculation of the radiosity solution; see [5, 20].

If one knows the solution of the radiosity equation and adds the glossy effects, one has the possibility of generating photo-realistic views into Ω, and one can study the appearance of rooms or the effect of different light sources before they are built; see [5, 20].

The radiosity equation is a transport equation, and it is mathematically a second kind of integral equation

$$(I - K)u(x) = E(x), \quad x \in S, \tag{1.1}$$

where the integral operator K is given by

$$(Ku)(x) := \frac{\rho(x)}{\pi} \int_S \beta(x,y) \frac{[\vec{n}_x \cdot (y-x)]\,[\vec{n}_y \cdot (x-y)]}{\|x-y\|^4}\, u(y)\, dy. \tag{1.2}$$

Here β is the visibility function, and ρ, $\rho(x) \in [0,1)$, $x \in S$, describes the reflectivity; see section 2. The vectors $\vec{n}_x$ and $\vec{n}_y$ are the inner normal vectors at point x, and y, $\vec{n}_y$ exist almost everywhere. We will shortly indicate the physical assumptions, which lead to (1.2). First, the radiosity $u : S \longrightarrow \mathbb{R}^+$, which has the physical unit $Watt/m^2$, does not depend on the wavelength, and so one has to solve this equation for different parts of the electromagnetic spectrum if this influence has to be considered. The outgoing radiance (see [12] for an exact definition of the radiance; here we will think

*Received by the editors September 14, 2000; accepted for publication (in revised form) June 8, 2001; published electronically October 11, 2001.
http://www.siam.org/journals/sima/33-3/37810.html

†Johannes Gutenberg–Universität Mainz, Mainz, Germany (hansen@mathematik.uni-mainz.de).

of a flux of radiative power, depending on the direction) at point $y \in S$ in the direction of $x \neq y$, $\vec{n}_y \cdot (x-y) > 0$ is given by

$$\frac{1}{\pi} \frac{[\vec{n}_y \cdot (x-y)]}{\|x-y\|} u(y) = \frac{1}{\pi} \cos(\angle(\vec{n}_y, x-y))\, u(y). \tag{1.3}$$

This very simple relation between the density of emitted power (radiosity) and the directed outgoing power density (radiance) is exactly fulfilled for black bodies. But in the theory of radiative transfer, this relation, known as the Lambertian cosine law, is often used as an approximation for nonblack bodies, so-called diffusive emitters. The factor $1/\pi$ in formula (1.3) is caused by an integration which relates the radiance with the radiosity; see [12, formula (2.8b)]. With respect to the distance r from y the radiance decays like

$$\frac{1}{r^2} \tag{1.4}$$

because of the conservation of energy. To get the power flux at the point x of the surface S we have to multiply the incoming radiance with the cosine of the angle between the surface normal and the direction of the incoming radiation

$$\cos(\angle(\vec{n}_x, y-x)) = \frac{[\vec{n}_x \cdot (y-x)]}{\|x-y\|} \tag{1.5}$$

to get the normal component of the flux. In our model the rays of light propagate along straight lines, so we have only to consider the visible part of the surface S if we calculate the total incoming flux at x

$$\int_S \beta(x,y) \frac{[\vec{n}_x \cdot (y-x)]}{\|x-y\|} \frac{1}{\|x-y\|^2} \frac{[\vec{n}_y \cdot (x-y)]}{\pi \|x-y\|} u(y) ds_y, \tag{1.6}$$

where we used (1.3)–(1.5). For the reflection we again consider the simplest possible case and assume that the fraction $\rho(x)$ of (1.6) is reflected. The reflected power should again be distributed according to the Lambertian cosine law (Lambertian diffuse reflection). So (1.1), written in the form

$$u(x) = E(x) + (Ku)(x), \tag{1.7}$$

states the fact that the radiosity (density of outgoing radiation) $u(x)$ at point x is caused by a source term $E(x)$ and reflected incoming radiation $(Ku)(x)$.

There are already a large number of publications on (1.1) in computer science; see the references in [5, 20]. The two dimensional case (cylindrical surfaces) (1.1) was analyzed by Atkinson [1], and Atkinson and Chandler [2] studied the three dimensional case. Rathsfeld [17] studied the behavior of the solution $\overline{u}$ of (1.1) in the three dimensional case along edges, and Qatanani [16] derived some mathematical properties of (1.1) in the case of smooth boundaries and studied different numerical methods for the two dimensional case. Investigations on different collocation methods for the three dimensional case and their performance can be found in the articles of Atkinson and Chien [3] and Atkinson, Chien, and Seol [4]. The Galerkin method is analyzed in the Master's thesis of Schon [19].

In the present article we follow closely the article of Elschner [8] on the double layer potential over polyhedral surfaces. But our direction is slightly different. Elschner's main result is the Fredholmness and the invertibility of the double layer operator on S. Our situation is different because the visibility function allows a very easy proof of the fact that the operator K of (1.2) is a contraction on $L^2(S)$; see also [16]. This already implies the invertibility of $I - K$, and Elschner's main result for our equation is proved. But we will use Elschner's Mellin transform techniques to study the local behavior of $\overline{u}$ near the vertices. On the other side, the visibility function introduces a number of regularity problems; these will be investigated in section 4.

In section 2 we formulate our problem and show the invertibility of $I - K$ in $L^2(S)$.

In section 3 we follow Elschner's analysis and study the equation and its solution locally near an arbitrary vertex of S. Our first main result is the invertibility of these local operators in a scale of weighted L^2-spaces; see Theorem 3.6. In Theorem 3.6 we describe in three steps a range of weighted L^2-spaces where the local operators are invertible. The explicit calculation of all suitable spaces seems to be impossible (see the formulas (3.32) and (3.34)) because the range depends on the zeros of some transcendental functions, but in the third part of this theorem it becomes clear that it is at least not an empty set. Figure 3 gives some impression of the fitting weights. The next important result in section 3 is the mapping property of the local operators between adjacent faces of S; see Lemmas 3.7–3.9. Again this result is very technical. But Corollary 3.11 shows an application for the case of a convex set Ω and a $\mathcal{C}^\infty$-function $E(x)$. In this case the regularity is close to the regularity results in [8] for the double layer potential. The reason is that for convex regions the visibility function $\beta(\cdot)$ is equal to one, and the local regularity results are not disturbed by nonlocal nonregular contributions. We hope that these two results will lead to error estimates for the boundary element method. The results of this section can also be used to derive the Fredholmness of (1.1), but this is already known from section 2. If the set Ω is not convex, one has to use discontinuity meshing in order to improve the convergence of the boundary element methods; see [10, 11, 9, 7]. But even then it is useful to know how the solution behaves near the lines of "discontinuity," and the present work is one step in this direction.

In section 3 we neglect the influence of the visibility function β. Now in section 4 we study its influence on the right-hand side of the localized equations of section 3. Even if the emissivity function is smooth on the faces of S, the right-hand sides of the local operators need not be smooth. This is a big difference from the case of the double layer potential and is caused by the shadow lines. By cutting down each face of S into smaller triangles we get a $\mathcal{C}^1$-regularity result, Lemma 4.1. In the proof of this lemma it becomes clear how these nonlocal effects create the nonregular contributions. Despite the fact that we have a lot of effects to consider, we try to summarize all contributions for a $\mathcal{C}^1$-function $E(x)$ in Corollary 4.2.

2. The radiosity equation and its localization near the vertices.

We consider a bounded domain $\Omega \subset \mathbb{R}^3$ with boundary $S := \partial\Omega$. The boundary S is assumed to be a polyhedron, and

$$S = \bigcup_{j=1}^{n} \Delta_j, \tag{2.1}$$

where Δ_j, $j = 1(1)n$, are closed triangles with

$$\dot{\Delta}_j \cap \dot{\Delta}_k = \emptyset, \quad j \neq k. \tag{2.2}$$

Here $\dot{\Delta}_j$ denotes the relative interior of Δ_j. We also use the following notation:

$$\dot{S} = \bigcup_{j=1}^{n} \dot{\Delta}_j. \tag{2.3}$$

For $x \in \dot{\Delta}_j$ the inner normal $\vec{n}(x)$ is well defined and by $\mathcal{E}_j$, $j = 1(1)n_E$, respectively, $\mathcal{V}_j$, $j = 1(1)n_V$, we denote the edges, respectively, the vertices of S. For $x \in \dot{\Delta}_j$ and $y \in \dot{\Delta}_k$ the visibility function $\beta(x, y)$ is given by

$$\beta(x,y) := \begin{cases} 1, & \{\lambda x + (1-\lambda)y \mid \lambda \in (0,1)\} \cap S = \emptyset \ \wedge \vec{n}(x) \cdot \vec{n}(y) < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

For the reflectivity function ρ we assume

$$\rho(x)|_{\Delta_j} = \rho_j \in [0, 1), \quad j = 1(1)n, \tag{2.5}$$

and at the beginning our assumption on the emissivity is

$$E \in L^2(S). \tag{2.6}$$

Now we can formulate the radiosity equation as

$$(I - K)u(x) = E(x), \qquad x \in S, \tag{2.7}$$

with the integral operator K given by

$$(Ku)(x) = \frac{\rho(x)}{\pi} \int_S \beta(x,y) \frac{[\vec{n}(x) \cdot (y-x)]\,[\vec{n}(y) \cdot (x-y)]}{\|x-y\|^4} u(y)dy. \tag{2.8}$$

Remark.

1. In the integration over S in (2.8) and in (2.7) we neglect the sets $\cup_{j=1}^{n_E} \mathcal{E}_j$ and $\cup_{j=1}^{n_V} \mathcal{V}_j$ which are of measure zero.

2. We can relax the assumptions on S. It is sufficient if S is the finite union of triangles (not necessary bounded) and on every triangle Δ_j there is a well defined normal $\vec{n}_j$. But we then also have to allow that pairs of triangles, for example Δ_j and Δ_{j+1}, are the same, because then we have to consider both sides and normals of every triangle.

A well-known fact for the radiosity equation is the following lemma.

LEMMA 2.1. *For $x \in \dot{S}$ we have*

$$(Ke)(x) = \rho(x) \ < \ 1, \tag{2.9}$$

where $e(x) = 1$, $x \in S$, is the unity function.

Proof. We will prove (2.9) for the case of a polyhedral surface S. The idea follows [5], and similar proofs can be found in [17] and [16].

Let $x \in \Delta_{j_0}$, and denote

$$\Delta_j(x) := \{y \in \Delta_j \mid \beta(x,y) = 1\},$$
$$S(x) := \bigcup_{j=1}^{n} \Delta_j(x).$$

We remark that each $\Delta_j(x)$ is a finite collection of smaller triangles because only the shadows of some other triangles are subtracted. This implies

$$S(x) = \bigcup_{j=1}^{\widetilde{n}} \delta_j(x),$$

with $\delta_j(x)$ a triangle for each j. The right-hand side of (2.8) can now be written in the following way:

$$\begin{aligned}
&\frac{\rho(x)}{\pi} \int_S \beta(x,y) \frac{[\vec{n}(x) \cdot (y-x)]\,[\vec{n}(y) \cdot (x-y)]}{\|x-y\|^4} dy \\
&= \frac{\rho(x)}{\pi} \int_{S(x)} \frac{[\vec{n}(x) \cdot (y-x)]\,[\vec{n}(y) \cdot (x-y)]}{\|x-y\|^4} dy \\
&= \frac{\rho(x)}{\pi} \sum_{j=1}^{\widetilde{n}} \int_{\delta_j(x)} \frac{[\vec{n}(x) \cdot (y-x)]\,[\vec{n}(y) \cdot (x-y)]}{\|x-y\|^4} dy.
\end{aligned}$$

Choose $\varepsilon > 0$ such that

$$\varepsilon < \min \left\{ \operatorname{dist}(x, \partial \Delta_{j_0}), \min_{j \neq j_0} \operatorname{dist}(x, \Delta_j) \right\}.$$

We assume that $x = 0$, $\Delta_{j_0} \subset \{(\xi_1, \xi_2, \xi_3) \,|\, \xi_3 = 0\}$, and we denote by

$$\partial B_\varepsilon^+(0) := \{(\xi_1, \xi_2, \xi_3) \,|\, \xi_1^2 + \xi_2^2 + \xi_3^2 = \varepsilon^2,\ \xi_3 \geq 0\}$$

the upper part of the sphere with radius ε. We further call $\widetilde{\delta}_j(x)$ the projection of $\delta_j(x)$ on $\partial B_\varepsilon^+(0)$,

$$\widetilde{\delta}_j(x) := \left\{ \varepsilon \frac{x}{\|x\|} \,|\, x \in \delta_j(x) \right\},$$

and $W_j(x)$ is the volume between these two surfaces (see Figure 1),

$$W_j(x) := \left\{ \lambda \frac{x}{\|x\|} \,|\, \lambda \in [\varepsilon, \|x\|],\ x \in \delta_j(x) \right\}.$$

We define

$$f(y) := \frac{\vec{n}(x) \cdot (y-x)}{\|x-y\|^4} (x-y)$$

and get by a short calculation

$$\operatorname{div}(f) = 0.$$

The Gaussian divergence theorem implies

$$\begin{aligned}
0 &= \int_{W_j(x)} \operatorname{div} f(y) dV(y) \\
&= \int_{\partial W_j(x)} f(y) \cdot \widetilde{\vec{n}}(y) dy \qquad \widetilde{\vec{n}} \text{ outer normal} \\
&= -\int_{\delta_j(x)} \frac{[\vec{n}(x) \cdot (y-x)]\,[\vec{n}(y) \cdot (x-y)]}{\|x-y\|^4} dy - \int_{\widetilde{\delta}_j(x)} \frac{[\vec{n}(x) \cdot (y-x)]\,[\vec{n}(y) \cdot (x-y)]}{\|x-y\|^4} dy.
\end{aligned}$$

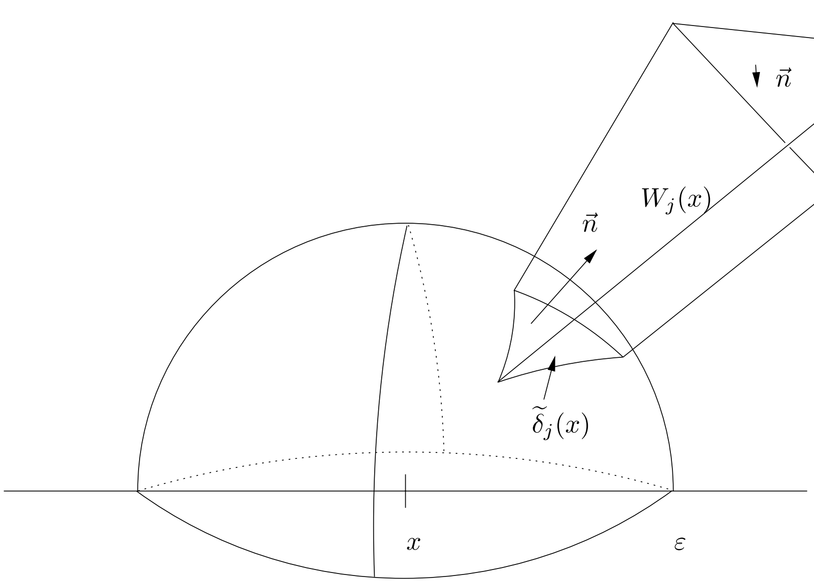


FIG. 1.

So we get

$$\int_{\delta_j(x)} \frac{[\vec{n}(x)\cdot(y-x)]\,[\vec{n}(y)\cdot(x-y)]}{\|x-y\|^4}dy = -\int_{\widetilde{\delta}_j(x)} \frac{[\vec{n}(x)\cdot(y-x)]\,[\vec{n}(y)\cdot(x-y)]}{\|x-y\|^4}dy$$

and finally

$$\begin{aligned} &\frac{\rho(x)}{\pi}\int_S \beta(x,y)\frac{[\vec{n}(x)\cdot(y-x)]\,[\vec{n}(y)\cdot(x-y)]}{\|x-y\|^4}dy \\ &= -\frac{\rho(x)}{\pi}\sum_{j=1}^{\widetilde{n}}\int_{\widetilde{\delta}_j(x)} \frac{[\vec{n}(x)\cdot(y-x)]\,[\vec{n}(y)\cdot(x-y)]}{\|x-y\|^4}dy \\ &= -\frac{\rho(x)}{\pi}\int_{\partial B_\varepsilon^+(0)} \frac{[\vec{n}(x)\cdot(y-x)]\,[\vec{n}(y)\cdot(x-y)]}{\|x-y\|^4}dy \\ &=: A(x), \end{aligned} \tag{2.10}$$

because the surface S is closed and in every direction of the upper half plane we meet the visible surface exactly one time. We neglect again sets of measure zero and remember $x = 0$, $\vec{n}(x) = (0,0,1)$, $\vec{n}(y) = y/\varepsilon$, $\|x-y\| = \varepsilon$. This gives us

$$\begin{aligned} A(x) &= -\frac{\rho(x)}{\pi}\int_0^{\pi/2}\int_0^{2\pi} \frac{\varepsilon\cos(\vartheta)(-1/\varepsilon)\varepsilon^2}{\varepsilon^4}\varepsilon\sin(\vartheta)d\varphi\,\varepsilon\,d\vartheta \\ &= \rho(x) \end{aligned}$$

and proves our result. □

Remark. The above proof shows that we get

$$(Ke)(x) \le \rho(x), \qquad x \in \dot{S}, \tag{2.11}$$

in the case in which S is not a closed surface. Then in (2.10) we get "$\leq$" instead of "$=$," because there may be a set of directions with measure greater than zero where x sees no part of S. One also has to consider the fact that the kernel function in (2.10) is positive.

We can copy the proof of [16, Lemma 2.3] and get the following theorem.

Theorem 2.2. *There exists a constant* $q_K < 1$,

$$q_K := \max_{j=1}^{n} \rho(x) \tag{2.12}$$

(see (2.5)*), such that*

$$\|K\|_{L^2(S)\to L^2(S)} \leq q_K. \tag{2.13}$$

The fixed point theorem of Banach shows the following corollary.

Corollary 2.3. *Equation* (2.7) *has exactly one solution* $\overline{u} \in L^2(S)$, *and for* $u_0 \in L^2(S)$ *the series* $(u_j)_{j\in\mathbb{N}}$ *given by*

$$u_j := E + Ku_{j-1}, \quad j \in \mathbb{N}, \tag{2.14}$$

converges linearly to $\overline{u}$.

We remark here that (2.14) is also of practical importance because it is used in computer graphics [5, 20] to solve (2.7) approximately. But then the function space $L^2(S)$ in (2.14) has to be replaced by some finite dimensional subspace.

The proper choice of these finite dimensional subspaces depends heavily on the smoothness of the solution $\overline{u}$. There are several reasons for the nonsmoothness of the solution $\overline{u}$. If the emissivity function is not smooth, along a line, for example, then $\overline{u}$ will also not be smooth along this line in general. This kind of singularity is very easy to handle because the right-hand side E of (2.7) is given and normally one knows its properties. The next reason is due to the visibility function; the solution along shadow lines may have some singularities. In section 4 we will study this effect in our context. Finally, along edges and near vertices, the smoothness of $\overline{u}$ is also not clear. The behavior of $\overline{u}$ along edges was studied in [17], and in the following section we will analyze the behavior of $\overline{u}$ near the vertices.

We choose one fixed vertex $\mathcal{V}_{j_0}$, $j_0 \in \{1, \dots, n_V\}$, and for simplicity we also assume

$$\mathcal{V}_{j_0} = 0. \tag{2.15}$$

For the localization we choose a function $\varphi \in \mathcal{C}^\infty(\mathbb{R}^3)$, $\varphi(x) \in [0,1]$, $x \in \mathbb{R}^3$, and

$$\varphi(x) = \begin{cases} 1, & \|x\| \leq \varepsilon, \\ 0, & \|x\| \geq 2\varepsilon, \end{cases} \tag{2.16}$$

where $\varepsilon > 0$ is so small that

$$\mathcal{V}_j \notin B_{3\varepsilon}(0), \quad j \neq j_0. \tag{2.17}$$

We get

$$\begin{aligned}[(1-\varphi+\varphi)\overline{u}](x) &= [(1-\varphi)+\varphi]K[[(1-\varphi)+\varphi]\overline{u}](x) + E(x) \\ &= [\varphi K\varphi\overline{u}](x) + [\varphi K(1-\varphi)\overline{u}](x) + [(1-\varphi)K\overline{u}](x) + E(x),\end{aligned}$$

and for $x \in B_\varepsilon(0)$ we have

$$(\varphi\overline{u})(x) = K(\varphi\overline{u})(x) + [K(1-\varphi)\overline{u}](x) + E(x). \tag{2.18}$$

From now on we assume, without loss of generality, that $\varepsilon = 1$; and by $\gamma = \gamma_{j_0}$ we denote

$$\gamma := S \cap \partial B_1(0), \tag{2.19}$$

with γ as a spherical polygon. The infinite polyhedral cone $\Gamma := \Gamma_{j_0}$ is defined by

$$\Gamma := \{r\omega \,|\, \omega \in \gamma,\ r \geq 0\}, \tag{2.20}$$

and the operator $\widetilde{K}$ on Γ is defined by

$$(\widetilde{K}u)(x) := \frac{\widetilde{\rho}(x)}{\pi} \int_\Gamma \beta(x,y) \frac{[\vec{n}(x)\cdot(y-x)]\,[\vec{n}(y)\cdot(x-y)]}{\|x-y\|^4} u(y)dy, \tag{2.21}$$

$u \in L^2(\Gamma)$ (in the next section we will see that $\widetilde{K}$ is well defined), $\widetilde{\rho}(x) := \rho(x/\|x\|)$, $x \neq 0$, and β is defined corresponding to (2.4).

By $\widetilde{u}$ and $\widetilde{E}$ we denote the extension by zero of the functions $\psi\overline{u}$ and ψE to Γ, where

$$\psi(x) := \varphi(3x), \tag{2.22}$$

$\text{supp}(\psi) \subset B_{2/3}(0)$. After some calculations we get

$$\begin{aligned}(I-\widetilde{K})\widetilde{u} &= \widetilde{E} + \psi K(1-\varphi)\overline{u} + (\psi\widetilde{K} - \widetilde{K}\psi)(\varphi\overline{u})\\ &=: f_1 + f_2 + f_3.\end{aligned} \tag{2.23}$$

We will study the equation

$$(I-\widetilde{K})v(x) = f(x), \quad x \in \Gamma, \tag{2.24}$$

in section 3, where we assume that f is sufficiently smooth. In section 4 we take a closer look at the functions f_2 and f_3 and recall some results of Rathsfeld [17].

3. The radiosity equation on an infinite polyhedral cone. Here we study (2.24), $\widetilde{K}$ given by (2.21), on the cone Γ of (2.20).

We assume that $(0,0,1)$ is a corner of the spherical polygon γ of (2.19), and we denote by F_1 and F_2 the two faces of Γ adjacent to edge $E_1 = \{(0,0,t)|t \geq 0\}$. After a suitable rotation, F_1 and F_2 have the following representation (see Figure 2):

$$\left\{\begin{aligned} F_1 &= \left\{ r\begin{pmatrix}\sin(\delta)\\ 0\\ \cos(\delta)\end{pmatrix} \,|\, \delta \in [0,\delta_1], \quad r \geq 0\right\},\\ F_2 &= \left\{ r\begin{pmatrix}\cos(\alpha)\sin(\delta)\\ \sin(\alpha)\sin(\delta)\\ \cos(\delta)\end{pmatrix} \,|\, \delta \in [0,\delta_2], \quad r \geq 0\right\}\end{aligned}\right. \tag{3.1}$$

with normals

$$\vec{n}_1 = \begin{pmatrix}0\\1\\0\end{pmatrix}, \quad \vec{n}_2 = \begin{pmatrix}\sin(\alpha)\\ -\cos(\alpha)\\ 0\end{pmatrix}. \tag{3.2}$$

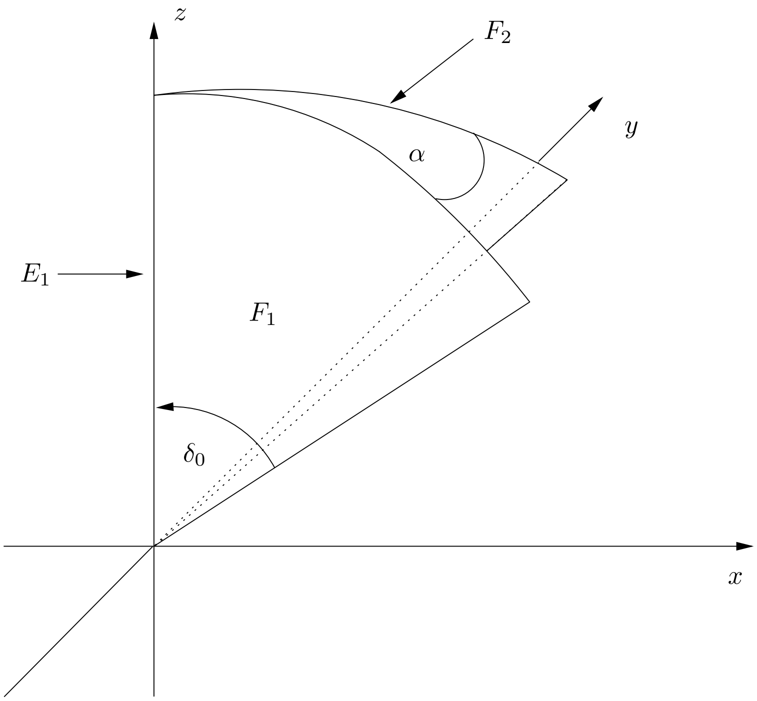


FIG. 2. *The two faces adjacent to edge E_1.*

Here $\alpha < \pi$ is the angle between F_1 and F_2. For $\alpha \geq \pi$ there would be no interaction between F_1 and F_2. We study the solution only local to the edge E_1, and so we assume $\delta_1 = \delta_2 = \delta_0 < \pi/2$. Further, δ_0 should be sufficiently small in order to guarantee that each part of F_1 can "see" each part of F_2. *This means that no occlusion occurs, and for the rest of this section we assume $\beta(x, y) = 1$.* Using the local coordinates from (3.1) and the notation $u_1 := v|_{F_1}$, respectively, $u_2 := v|_{F_2}$, we can rewrite (2.24) as

$$\text{(3.3)} \qquad \left\{ \begin{array}{lcl} u_1(r,\delta) - \dfrac{\rho_1}{\pi} \displaystyle\int_0^\infty \int_0^{\delta_0} k_1(r,\delta,r',\delta') u_2(r',\delta') d\delta' \, dr' & = & f_1(r,\delta), \\ u_2(r,\delta) - \dfrac{\rho_2}{\pi} \displaystyle\int_0^\infty \int_0^{\delta_0} k_1(r,\delta,r',\delta') u_1(r',\delta') d\delta' \, dr' & = & f_2(r,\delta) \end{array} \right.$$

with k_1 given by

$$\text{(3.4)} \qquad k_1(r,\delta,r',\delta') := \frac{\sin(\alpha)^2 \sin(\delta) \sin(\delta') r r'^2}{(r'^2 - 2\omega(\delta,\delta') r r' + r^2)^2},$$

and

$$\text{(3.5)} \qquad \omega(\delta,\delta') := \cos(\alpha)\sin(\delta)\sin(\delta') + \cos(\delta)\cos(\delta')$$

is the cosine of the angle between

$$\begin{pmatrix} \sin(\delta) \\ 0 \\ \cos(\delta) \end{pmatrix} \text{ and } \begin{pmatrix} \cos(\alpha)\sin(\delta') \\ \sin(\alpha)\sin(\delta') \\ \cos(\delta') \end{pmatrix}.$$

We also assume that the reflectivity functions ρ_1 and ρ_2 are constant on F_1, respectively, F_2. We will later often use the fact

$$\omega(\delta, \delta') \in [0, 1], \quad \delta \in [0, \delta_0], \tag{3.6}$$

where we used $\delta_0 < \pi/2$. We remark that the functions f_1 and f_2 also include the contributions of the integral over other parts of Γ. Using the new functions

$$v_1 := \frac{u_1 - \sqrt{\rho_1/\rho_2}u_2}{2} \text{ and } v_2 := \frac{u_1 + \sqrt{\rho_1/\rho_2}u_2}{2} \tag{3.7}$$

and multiplying (3.3) by the invertible matrix

$$\frac{1}{2}\begin{pmatrix} 1 & -\sqrt{\rho_1/\rho_2} \\ 1 & \sqrt{\rho_1/\rho_2} \end{pmatrix},$$

we see that (3.3) is equivalent to

$$\left.\begin{array}{rcl} v_1 - \overline{K}v_1 & = & \frac{1}{2}(f_1 - \sqrt{\rho_1/\rho_2}f_2) \\ v_2 - \overline{K}v_2 & = & \frac{1}{2}(f_1 + \sqrt{\rho_1/\rho_2}f_2) \end{array}\right\}, \tag{3.8}$$

where the integral operator $\overline{K}$ on $[0, \infty) \times [0, \delta_0]$ is given by

$$(\overline{K}v)(r, \delta) := \int_0^\infty \int_0^{\delta_0} \overline{k}(r, \delta, r', \delta')v(r', \delta')d\delta'\, dr', \tag{3.9}$$

and the kernel

$$\begin{aligned} \overline{k}(r, \delta, r', \delta') &:= \frac{\sqrt{\rho_1\rho_2}}{\pi} \frac{\sin(\alpha)^2 \sin(\delta) \sin(\delta')rr'^2}{(r'^2 - 2\omega(\delta, \delta')rr' + r^2)^2} \qquad (3.10) \\ &= \frac{\sqrt{\rho_1\rho_2}}{\pi} \sin(\alpha)^2 \frac{\sin(\delta) \sin(\delta')(r/r')}{((r/r')^2 - 2\omega(\delta, \delta')(r/r') + 1)^2} \frac{1}{r'} \\ &=: \overline{k}_1(\delta, \delta', r/r') \, \frac{1}{r'}. \qquad (3.11) \end{aligned}$$

For suitable v (for example, $v \in \mathcal{C}_0^\infty$) we can now write the integral operator $\overline{K}$ as a Mellin convolution operator with an operator valued kernel:

$$(\overline{K}v)(r, \delta) = \int_0^\infty \left(\int_0^{\delta_0} \overline{k}_1(\delta, \delta', r/r')v(r', \delta')d\delta' \right) \frac{dr'}{r'}. \tag{3.12}$$

Because the Mellin convolution is diagonalized by the Mellin transform $\mathcal{M}$, it is now natural to apply the Mellin transform with respect to r to the operator $\overline{K}$.

We summarize the definition and some properties of the Mellin transform $\mathcal{M}$:

$$(\mathcal{M}v)(z) := \int_0^\infty t^{z-1}v(t)\, dt; \tag{3.13}$$

see [13]. If we introduce the weighted L^2-space $L^2_\nu([0, \infty))$ with

$$\|v\|_{L^2_\nu} := \left(\int_0^\infty \left|t^{-\nu}v(t)\right|^2 dt \right)^{1/2}, \tag{3.14}$$

then we get that the function

$$x \longrightarrow (\mathcal{M}v)\left(\frac{1}{2} - \nu + ix\right), \quad v \in L^2_\nu([0,\infty)), \tag{3.15}$$

is well defined and $(\mathcal{M}v)(1/2 - \nu + i\cdot) \in L^2(\mathbb{R})$. We denote the mapping

$$v(x) \longrightarrow (\mathcal{M}v)\left(\frac{1}{2} - \nu + ix\right), \quad x \in \mathbb{R}, \tag{3.16}$$

by $\mathcal{M}_{Re(z)=1/2-\nu}$ and keep in mind

$$\mathcal{M}_{Re(z)=1/2-\nu} : \; L^2_\nu([0,\infty)) \xrightarrow{1:1} L^2(\mathbb{R}); \tag{3.17}$$

see [13, Theorem 3.2], where the restriction $\nu \le 1/2$ is caused by his definition of his set $\mathcal{E}$. We apply this Mellin transform to the operator $\overline{K}$ with respect to r and get

$$\begin{aligned} \mathcal{M}(\overline{K}v)(z,\delta) \; &= \int_0^{\delta_0} [\mathcal{M}\overline{k}_1(\delta,\delta',\cdot)](z)[\mathcal{M}v](z,\delta')d\delta' \\ &=: [\mathcal{A}(z)(\mathcal{M}v)(z)](\delta). \end{aligned} \tag{3.18}$$

This means that we get for all z in some area of the complex domain an integral operator $\mathcal{A}(z)$ on the interval $[0,\delta_0]$ with kernel

$$\left\{ \begin{array}{lcl} \overline{k}(z,\delta,\delta') & := & \dfrac{\sqrt{\rho_1\rho_2}}{\pi}\sin(\alpha)^2\sin(\delta)\sin(\delta')\kappa(z,\delta,\delta'), \\ \kappa(z,\delta,\delta') & := & (1-\omega(\delta,\delta'))^{-3/2}\kappa_1(z,\delta,\delta'), \\ \kappa_1(z,\delta,\delta') & := & 2^{3/2}B(1+z,3-z)\,{}_2F_1\left(\dfrac{3}{2}-z, z-\dfrac{1}{2}, \dfrac{5}{2}, \dfrac{1+\omega(\delta,\delta')}{2}\right). \end{array} \right. \tag{3.19}$$

Here B is the Beta function, and ${}_2F_1$ is the hypergeometric function; see [15]. Formula (3.19) follows from the integral table in [14, p. 310, formula 22], where the exponent $\nu - 0.5$ is corrected to $0.5 - \nu$. The above formula is correct for all $z \in \Sigma_{-1,3}$,

$$\Sigma_{\alpha,\beta} := \{z \in \mathbb{C} |\, \alpha < \mathrm{Re}(z) < \beta\}, \quad \alpha < \beta, \quad \alpha,\beta \in \mathbb{R}. \tag{3.20}$$

To analyze this operator we first estimate the function κ_1.

LEMMA 3.1. *The function $\kappa_1(z,\delta,\delta')$ is holomorphic in the strip $\Sigma_{-1,3}$. We have*

$$|\kappa_1(z,\delta,\delta')| \le \kappa_1(\mathrm{Re}(z),\delta,\delta'),$$

and for $s \in [-0.5, 2.5]$

$$0 \;\le\; \kappa_1(s,\delta,\delta') \le \left\{ \begin{array}{ll} \frac{\sqrt{2}}{8}\pi, & s \in [0.5, 1.5], \\ \frac{\sqrt{2}}{3}\Gamma(s+1)\Gamma(3-s), & s \in [-0.5, 0.5] \cup [1.5, 2.5]. \end{array} \right.$$

Proof. We have

$$\kappa(z,\delta,\delta') = \int_0^\infty \frac{t^{z-1}t}{(1-2\omega(\delta,\delta')+t^2)^2}dt.$$

This implies the following.

1. The function $\kappa(z)$ is holomorphic in $\Sigma_{-1,3}$ because of the convergence of the integral in this domain.

2.

$$\begin{aligned}|\kappa_1(z,\delta,\delta')| &\le (1-\omega(\delta,\delta'))^{3/2}\int_0^\infty \frac{|t^z|}{(1-2\omega(\delta,\delta')t+t^2)^2}dt\\ &= (1-\omega(\delta,\delta'))^{3/2}\int_0^\infty \frac{t^{\mathrm{Re}(z)}}{(1-2\omega(\delta,\delta')t+t^2)^2}dt\\ &= \kappa_1(\mathrm{Re}(z),\delta,\delta').\end{aligned}$$

For all $z \in \mathbb{C}$ we have

$$\mathrm{Re}\left(\frac{5}{2}-\left(\frac{3}{2}-z\right)-\left(z-\frac{1}{2}\right)\right) = \mathrm{Re}\left(\frac{3}{2}\right) \ > \ 0.$$

It is known from [21] that the following limit exists:

$$\lim_{w\to 1-} {}_2F_1(3/2-z, z-1/2, 5/2, w) = \frac{\Gamma(5/2)\Gamma(3/2)}{\Gamma(1+z)\Gamma(3-z)}.$$

By definition (see also [21])

$${}_2F_1(3/2-z,z-1/2,5/2,w) = \sum_{j=0}^\infty \underbrace{\frac{(3/2-z)_j(z-1/2)_j}{(5/2)_j j!}}_{=:a_j(z)} w^j.$$

If $s \in [1/2, 3/2]$, we get $a_j(s) \ge 0$. This implies that ${}_2F_1(3/2-s, s-1/2, 5/2, w)$ is monotone increasing with respect to w. By (3.6) we know

$$\frac{1+\omega(\delta,\delta')}{2} \in [1/2,1] \quad \forall \delta,\delta'.$$

So we estimate

$$\begin{aligned}\kappa_1(s,\delta,\delta') &\le 2^{3/2}B(1+s,3-s)_2F_1(3/2-s,s-1/2,5/2,1)\\ &= 2^{3/2}\frac{\Gamma(1+s)\Gamma(3-s)}{\Gamma(4)}\frac{\Gamma(5/2)\Gamma(3/2)}{\Gamma(1+s)\Gamma(3-s)}\\ &= 2^{3/2}\frac{1}{3!}\frac{3}{2}\Gamma(3/2)^2\\ &= \sqrt{2}\frac{\pi}{8}.\end{aligned}$$

If $s \in [-1/2, 1/2]$, we have

$$\begin{aligned}(s-1/2)_j &\le 0, \quad j \ge 1,\\ (3/2-s)_j &> 0, \quad j \ge 1,\\ \Longrightarrow a_j(s) &\le 0, \quad j \ge 1.\end{aligned}$$

So ${}_2F_1(3/2-s, s-1/2, 5/2, w)$ is monotone decreasing in w, and finally,

$$\begin{aligned}\kappa_1(s,\delta,\delta') &\le 2^{3/2}B(1+s,3-s)_2F_1(3/2-s,s-1/2,5/2,0)\\ &= 2^{3/2}\frac{\Gamma(1+s)\Gamma(3-s)}{\Gamma(4)}\,1\\ &= \frac{\sqrt{2}}{3}\Gamma(1+s)\Gamma(3-s).\end{aligned}$$

If $s \in [3/2, 5/2]$, it follows that

$$\begin{aligned}(s-1/2)_j &\geq 0, \quad j \geq 1,\\ (3/2-s)_j &\leq 0, \quad j \geq 1,\end{aligned}$$

and, similar to the above, we can estimate

$$\kappa_1(s,\delta,\delta') \leq \frac{\sqrt{2}}{3}\Gamma(1+s)\Gamma(3-s).$$

Now we have proved our estimates. □

In the proof we saw that it is sufficient to estimate ${}_2F_1(3/2-z, z-1/2, 5/2, w)$ in the vicinity of $w=1$. This proves the following corollary.

Corollary 3.2. *Let $\varepsilon, \varepsilon' > 0$. Then we get*

$$|\kappa_1(z,\delta,\delta')| \leq \frac{\sqrt{2}}{8}\pi + \varepsilon', \tag{3.21}$$

$z \in [-1+\varepsilon, 3-\varepsilon]$, $\delta, \delta' \in [0, \delta_0(\varepsilon,\varepsilon')]$, if $\delta_0(\varepsilon,\varepsilon')$ is sufficiently small.

Our next step is the study of the z-independent part of the integral operator $\mathcal{A}(z)$. We denote this operator by $\mathcal{B}$:

$$(\mathcal{B}v)(\delta) := \frac{\sqrt{\rho_1\rho_2}}{\pi}\sin(\alpha)^2\int_0^{\delta_0}\frac{\sin(\delta)\sin(\delta')}{(1-\omega(\delta,\delta'))^{3/2}}v(\delta')d\delta'. \tag{3.22}$$

The next lemma shows that $\mathcal{B}$ is closely connected to a Mellin operator (see also the proof of Theorem 2.1 in [8]).

Lemma 3.3. *We have*

$$\mathcal{B} = T_2^{-1} \circ T_1^{-1} \circ \mathcal{B}_2 \circ T_1 \circ T_2,$$

where $\mathcal{B}_2$ is a finite Mellin convolution on $[0, \vartheta_0]$, $\vartheta_0 := \tan(\delta_0/2)$,

$$(\mathcal{B}_2 w)(\vartheta) := \int_0^{\vartheta_0} l_2(\vartheta/\vartheta')w(\vartheta')\frac{d\vartheta'}{\vartheta'}, \tag{3.23}$$

and

$$l_2(t) := 2^{3/2}\frac{\sqrt{\rho_1\rho_2}}{\pi}\sin(\alpha)^2\frac{t}{(1-2\cos(\alpha)t+t^2)^{3/2}}. \tag{3.24}$$

The invertible operators T_1 and T_2 are given by

$$T_1 : L^2_\nu([0,\delta_0]) \longrightarrow L^2_\nu([0,\delta_0]), (T_1 v)(\delta) := v(\delta)/\cos(\delta/2) \tag{3.25}$$

and

$$T_2 : L^2_\nu([0,\delta_0]) \longrightarrow L^2_\nu([0,\vartheta_0]), (T_2 v)(\vartheta) := v(2\text{arctan}(\vartheta)). \tag{3.26}$$

Proof. In formula (3.22) we use the addition theorems for sine and cosine and get the following expressions for the numerator, respectively, the denominator, under the integral sign:

$$\begin{aligned}\sin(\delta)\sin(\delta') &= 4\sin(\delta/2)\sin(\delta'/2)\cos(\delta/2)\cos(\delta'/2),\\ 1-\omega(\delta,\delta') &= 2\cos(\delta/2)^2\sin(\delta'/2)^2 + 2\sin(\delta/2)^2\cos(\delta'/2)^2\\ &\quad - 4\cos(\alpha)\sin(\delta/2)\sin(\delta'/2)\cos(\delta/2)\cos(\delta'/2).\end{aligned}$$

Substituting this in (3.22) gives

$$\begin{aligned}
&\mathcal{B}w(\delta)\\
&= \sqrt{2}\frac{\sqrt{\rho_1\rho_2}}{\pi}\sin(\alpha)^2\frac{1}{\cos(\delta/2)}\int_0^{\delta_0}\frac{\tan(\delta/2)\tan(\delta'/2)}{(\tan(\delta/2)^2+\tan(\delta'/2)^2-2\cos(\alpha)\tan(\delta/2)\tan(\delta'/2))^{3/2}}\\
&\qquad\qquad\times\cos(\delta'/2)w(\delta')\frac{d\delta'}{\cos(\delta'/2)^2}\\
&= (T_1^{-1}\circ\mathcal{B}_1\circ T_1)w(\delta)
\end{aligned}$$

with T_1 defined in (3.25) and $\mathcal{B}_1$ defined according to

$$\begin{aligned}
&(\mathcal{B})_1 w(\delta)\\
&\quad := \sqrt{2}\frac{\sqrt{\rho_1\rho_2}}{\pi}\sin(\alpha)^2\int_0^{\delta_0}\frac{\tan(\delta/2)\tan(\delta'/2)}{(\tan(\delta/2)^2+\tan(\delta'/2)^2-2\cos(\alpha)\tan(\delta/2)\tan(\delta'/2))^{3/2}}\\
&\qquad\qquad\times w(\delta')\frac{d\delta'}{\cos(\delta'/2)^2}.
\end{aligned}$$

Substituting $\vartheta := \tan(\delta/2)$, respectively, $\vartheta' := \tan(\delta'/2)$, which implies

$$d\vartheta' = \frac{1}{2}\frac{1}{\cos(\delta'/2)^2}d\delta',$$

shows

$$\begin{aligned}
&\mathcal{B}_1 w(2\mathrm{arctan}(\vartheta))\\
&\quad = 2^{3/2}\frac{\sqrt{\rho_1\rho_2}}{\pi}\sin(\alpha)^2\int_0^{\vartheta_0}\frac{\vartheta\vartheta'}{(\vartheta^2+\vartheta'^2-2\cos(\alpha)\vartheta\vartheta')^{3/2}}w(2\mathrm{arctan}(\vartheta'))d\vartheta'\\
&\quad = 2^{3/2}\frac{\sqrt{\rho_1\rho_2}}{\pi}\sin(\alpha)^2\int_0^{\vartheta_0}\frac{\vartheta/\vartheta'}{((\vartheta/\vartheta')^2+1-2\cos(\alpha)\vartheta/\vartheta')^{3/2}}w(2\mathrm{arctan}(\vartheta'))\frac{d\vartheta'}{\vartheta'}.
\end{aligned}$$

The last equation implies

$$T_2\circ\mathcal{B}_1 = \mathcal{B}_2\circ T_2.$$

This proves the representation formula for $\mathcal{B}$, formula (3.23), and the invertibility of T_1 and T_2 is clear because of $\delta_0 < \pi/2$. □

Now we will give estimates for the operators in Lemma 3.3 to finally get a bound for the operator norm of $\mathcal{A}(z)$.

LEMMA 3.4. *The operators T_1 and T_2 and their inverses are continuous and fulfill the following estimates:*

$$(3.27)\quad \begin{cases} \|T_1\| \;\leq\; \dfrac{1}{\cos(\delta_0/2)}, & \|T_2\| \;\leq\; \dfrac{1}{\sqrt{2}\cos(\delta_0/2)}\max\limits_{\delta\in[0,\delta_0]}\left(\dfrac{\tan(\delta/2)}{\delta}\right)^{-\nu},\\[2ex] \|T_1^{-1}\| \;\leq\; 1, & \|T_2^{-1}\| \;\leq\; \sqrt{\dfrac{2^{1-2\nu}}{1+\vartheta_0^2}\max\limits_{\vartheta\in[0,\vartheta_0]}\left(\dfrac{\mathrm{arctan}(\vartheta)}{\vartheta}\right)^{-\nu}}. \end{cases}$$

Here we omit the domain of definition and the range of the operators.

Proof. The calculation of all of the above bounds is straightforward, and we give only the proof for $\|T_2\|_{L^2_\nu}$.

$$\begin{aligned}\|T_2 v\|_{L^2_\nu([0,\delta_0])^2\to L^2_\nu([0,\vartheta_0])} &= \int_0^{\vartheta_0} \vartheta^{-2\nu} v(2\mathrm{arctan}(\vartheta))\,d\vartheta \\ &= \int_0^{\delta_0} \tan(\delta/2)^{-2\nu} v(\delta)^2 \frac{d\delta}{2\cos(\delta/2)^2} \\ &= \int_0^{\delta_0} \delta^{-2\nu} v(\delta)^2 \underbrace{\left(\frac{1}{2\cos(\delta/2)^2}\left(\frac{\tan(\delta/2)}{\delta}\right)^{-2\nu}\right)}_{(*)} d\delta.\end{aligned}$$

The maximum over $\delta \in [0,\delta_0]$ of the $(*)$-term proves the bound for T_2. □

In the following we denote the product of all four bounds by $\sigma(\delta_0)$,

$$\sigma(\delta_0) := \|T_1\|_{L^2_\nu([0,\delta_0])}\|T_1^{-1}\|_{L^2_\nu([0,\delta_0])}\|T_2\|_{L^2_\nu([0,\delta_0])\to L^2_\nu([0,\vartheta_0])}\|T_2^{-1}\|_{L^2_\nu([0,\vartheta_0])\to L^2_\nu([0,\delta_0])}, \tag{3.28}$$

and it is easy to see that

$$\lim_{\delta_0\to 0}\sigma(\delta_0) = 1. \tag{3.29}$$

An upper bound for the norm of the operator $\mathcal{B}_2$ is given in the next lemma.

LEMMA 3.5. *The Mellin operator $\mathcal{B}_2$ is a bounded operator in $L^2_\nu([0,\infty))$, $\nu \in (-3/2, 3/2)$. We get*

$$\begin{aligned}&\|\mathcal{B}_2\|_{L^2_\nu} \\ &\quad\le 2^{5/2}\frac{\sqrt{\rho_1\rho_2}}{\pi}(1+\cos(\alpha))B(3/2-\nu, 3/2+\nu)\,{}_2F_1\left(1/2+\nu, 1/2-\nu, 2, \frac{1+\cos(\alpha)}{2}\right).\end{aligned} \tag{3.30}$$

This can be estimated further by

$$\begin{aligned}&\|\mathcal{B}_2\|_{L^2_\nu} \\ &\quad\le 2^{3/2}\frac{\sqrt{\rho_1\rho_2}}{\pi}(1+\cos(\alpha))\left\{\begin{array}{ll} 1, & \nu \in [-0.5, 0.5], \\ \Gamma(3/2-\nu)\Gamma(3/2+\nu), & \nu \in (-1.5, -.5)\cup(0.5, 1.5).\end{array}\right.\end{aligned} \tag{3.31}$$

Proof. The Mellin transform $\mathcal{M}_{Re(z)=1/2-\nu}$ transforms the operator $\mathcal{B}_2$ on L^2_ν into a multiplication operator with

$$\begin{aligned}\widehat{l}_2(w) = 2^{3/2}\frac{\sqrt{\rho_1\rho_2}}{\pi}\sin^2(\alpha)\,2B(w+1, 2-w)\frac{1}{1-\cos(\alpha)}\,{}_2F_1\left(1-w, w, 2, \frac{1+\cos(\alpha)}{2}\right),\\ \mathrm{Re}(w) = 1/2-\nu;\end{aligned}$$

see [14, p. 310, formula 22]. This transform exists for $w \in \Sigma_{-1,2}$. Because the function l_2 is positive, we again have the relation

$$\left|\widehat{l}_2(w)\right| \le \widehat{l}_2(\mathrm{Re}(w)).$$

This, together with the range for w, proves formula (3.30) because a multiplication operator is bounded by its essential maximum. For $\nu \in [-0.5, 0.5]$ the coefficients of the power series of ${}_2F_1$ are positive, so we get ($w = 1/2 - \nu$)

$$\begin{aligned} {}_2F_1\left(1/2+\nu, 1/2-\nu, 2, \frac{1+\cos(\alpha)}{2}\right) &\leq {}_2F_1(1/2+\nu, 1/2-\nu, 2, 1) \\ &= \frac{\Gamma(2)\Gamma(1)}{\Gamma(3/2-\nu)\Gamma(3/2+\nu)}. \end{aligned}$$

Together with

$$B(3/2-\nu, 3/2+\nu) = \frac{\Gamma(3/2-\nu)\Gamma(3/2+\nu)}{\Gamma(3)},$$

we get the upper bound in (3.31). For $\nu \in (-1.5, 0.5]$, respectively, $\nu \in [0.5, 1.5)$, the coefficients of the power series for ${}_2F_1(1/2+\nu, 1/2-\nu, 2, w)$ are negative (with the exception of the coefficient of w^0). This proves

$$\begin{aligned} {}_2F_1\left(1/2+\nu, 1/2-\nu, 2, \frac{1+\cos(\alpha)}{2}\right) &\leq {}_2F_1(1/2+\nu, 1/2-\nu, 2, 0) \\ &= 1, \end{aligned}$$

and now the second estimate in (3.31) is also proved. □

Theorem 3.6. *The operator $\overline{K}$ (see* (3.9)*) is a continuous operator on*

$$L^2_{\nu_1}([0,\infty)) \otimes L^2_{\nu_2}([0,\delta_0]), \quad \nu_1 \in (-2.5, 1.5),\ \nu_2 \in (-1.5, 1.5).$$

1. *The operator $(I - \overline{K})$ is invertible if*

$$B_0(\nu_1,\nu_2)\sigma_0(\delta_0) < 1 \tag{3.32}$$

holds, where

(3.33)$B_0(\nu_1,\nu_2)$

$$:= \sqrt{\rho_1\rho_2}(1+\cos(\alpha))B(3/2-\nu_2, 3/2+\nu_2){}_2F_1\left(1/2+\nu_2, 1/2-\nu_2, 2, \frac{1+\cos(\alpha)}{2}\right)$$
$$\times \begin{cases} 1, & (\nu_1,\nu_2) \in [-1,0] \times (-1.5, 1.5), \\ \frac{8}{3}\Gamma(3/2-\nu_1)\Gamma(5/2+\nu_1), & (\nu_1,\nu_2) \in [(-2,-1) \cup (0,1)] \times (-1.5, 1.5). \end{cases}$$

2. *$(I - \overline{K})$ is invertible if*

(3.34)$B_1(\nu_2)$

$$:= \sqrt{\rho_1\rho_2}(1+\cos(\alpha))B(3/2-\nu_2, 3/2+\nu_2){}_2F_1\left(1/2+\nu_2, 1/2-\nu_2, 2, \frac{1+\cos(\alpha)}{2}\right)$$
$$< 1,$$

$(\nu_1,\nu_2) \in (-2.5, 1.5) \times (-1.5, 1.5)$, *if δ_0 is sufficiently small.*

3. *If δ_0 is sufficiently small, the operator $I - \overline{K}$ is invertible on*

$$L^2_{\nu_1}([0,\infty)) \otimes L^2_{\nu_2}([0,\delta_0]), (\nu_1,\nu_2) \in [-1,0] \times [-0.5, 0.5]. \tag{3.35}$$

Proof.

1. After the application of the Mellin transform $\mathcal{M}_{Re(z)=1/2-\nu_1}$, the operator $\overline{K}$ is transformed into the parameter-dependent operator $\mathcal{A}(1/2-\nu_1+i\tau)$, $\tau \in \mathbb{R}$. For this operator we have

$$\begin{aligned}
&|[[\mathcal{A}(1/2-\nu_1+i\tau)]u](\delta)| \\
&\quad \le \frac{\sqrt{\rho_1\rho_2}}{\pi}\sin(\alpha)^2\int_0^{\delta_0}\frac{\sin(\delta)\sin(\delta')}{(1-\omega(\delta,\delta'))^{3/2}}\,|\kappa_1(1/2-\nu_1+i\tau,\delta,\delta')|\,|w(\delta')|\,d\delta' \\
&\quad \le C(\nu_1)[\mathcal{B}_2\,|w|](\delta).
\end{aligned}$$

Now using the bounds for $C(\nu_1)$, given in Lemmas 3.1 and 3.5 together with Lemmas 3.3 and 3.4, we get the above bound for $\mathcal{A}(1/2-\nu_1+i\tau)$ on $L^2_{\nu_2}$, independent of τ. This proves our result.

2. The proof of part 2 follows from Corollary 3.2.

3. For (ν_1,ν_2) in the above range we get by Lemmas 3.1 and 3.5 the bound

$$\begin{aligned}
B_0(\nu_1,\nu_2) &\le \frac{\sqrt{2}}{8}\pi \times \frac{\sqrt{\rho_1\rho_2}}{\pi}2^{3/2}(1+\cos(\alpha)) \\
&= \sqrt{\rho_1\rho_2}\frac{1+\cos(\alpha)}{2} \\
&< 1.
\end{aligned}$$

By (3.29) we know that we can choose δ_0 sufficiently small to guarantee

$$B_0(\nu_1,\nu_2)\sigma(\delta_0) < 1,$$

which proves part 3. □

Remark. The result in part 3 of the above theorem shows that the set of values (ν_1,ν_2) in parts 1 and 2 for which $(I-\overline{K})$ is an invertible operator is nonempty. The author knows no explicit formula which describes the boundary of the (ν_1,ν_2)-area, in dependence on α, where the operator $(I-\overline{K})$ is invertible. But the function B_1 in (3.34) depends only on ν_2, and one can calculate *numerically* $\overline{\nu}_2(\alpha)$ with $B_1(\overline{\nu}_2(\alpha)) = 1$. Because of the symmetry of B_1, this implies that

$$\begin{aligned}
(I-\overline{K}):\ L^2_{\nu_1}([0,\infty))\otimes L^2_{\nu_2}([0,\delta_0]) &\xrightarrow{1:1} L^2_{\nu_1}([0,\infty))\otimes L^2_{\nu_2}([0,\delta_0]), \\
&(\nu_1,\nu_2)\in(-2.5,1.5)\times(-\overline{\nu}_2(\alpha),\overline{\nu}_2(\alpha))
\end{aligned}$$

if δ_0 is small enough.

Here it is interesting that the range of suitable ν_2 values increases near $\alpha = 0$ (at least for the more realistic values $\sqrt{\rho_1\rho_2} < 1$) and has a minimum for some positive angle $\overline{\alpha} = \overline{\alpha}(\sqrt{\rho_1\rho_2})$. This is *different* than in the case of the classical double layer potential, where the singularities are getting stronger when the angle goes to zero. Here Figure 3 suggests that there exists some worst angle for the regularity, which is greater than zero. To the author this property is not clear and cannot be explained at the moment. But we would like to mention that this behavior can also be seen in Table 1 of [17], although Rathsfeld has not explicitly mentioned this phenomenon.

Now we want to study the smoothing properties of the operator $\overline{K}$. It is our aim to study the behavior of

$$\sin(\delta)^{m+k}\left(r\frac{\partial}{\partial r}\right)^k\left(\frac{\partial}{\partial\delta}\right)^m\overline{k}_1\left(\delta,\delta',\frac{r}{r'}\right); \tag{3.36}$$

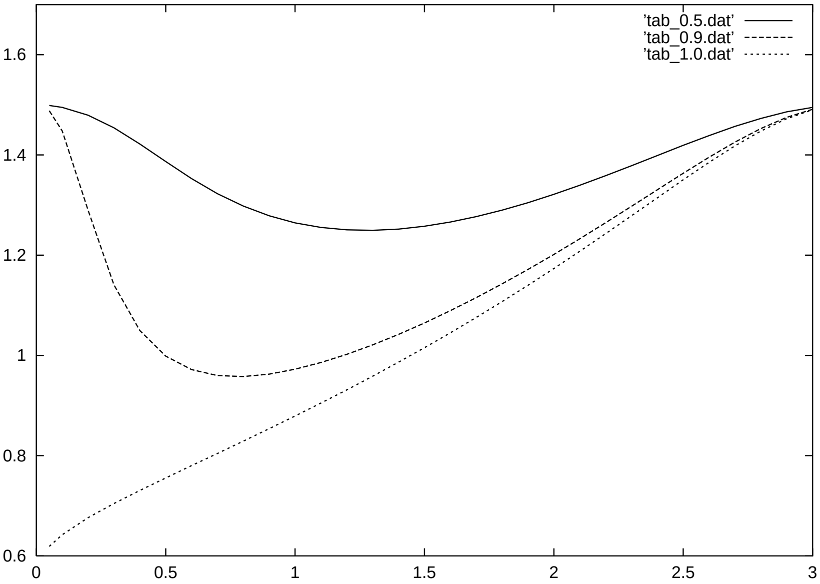


FIG. 3. *The curves* $\overline{\nu}_2(\alpha)$ *for three different values of* $\sqrt{\rho_1\rho_2}$: 0.5, 0.9, *and* 1.0.

see (3.11). Instead of $r\frac{\partial}{\partial r}$, we will write $t\frac{\partial}{\partial t}$ and consider $\overline{k}_1(\delta,\delta',t)$. We recall here that the function $\overline{k}_1$ is equivalent to k_1; see (3.4). We will see that an integral operator with kernel (3.36) behaves like the operator $\overline{K}$. But first we have to prove two technical lemmas. Instead of $\overline{k}_1$ we will use the equivalent but simpler kernel

$$\overline{k}_2(\delta,\delta',t) := \frac{\sin(\delta)t}{(1-2\omega(\delta,\delta')t+t^2)^2}. \tag{3.37}$$

LEMMA 3.7. *Let* $m \in \mathbb{N}_0$. *Then*

$$\begin{aligned}
&\left(\frac{\partial}{\partial\delta}\right)^m \overline{k}_2(\delta,\delta',t)\\
&= \sum_{n=0,2|n}^{m} \begin{pmatrix} m \\ n \end{pmatrix} \sin^{(m-n)}(\delta) \left(\sum_{j=1}^{n/2} \frac{t^{j+1} f_{n,j}}{(1-2\omega t+t^2)^{2+j}} + \sum_{j=n/2+1}^{n} \frac{t^{j+1}\omega'^{2j-n} f_{n,j}}{(1-2\omega t+t^2)^{2+j}} \right)\\
&\quad + \sum_{n=0,2\nmid n}^{m} \begin{pmatrix} m \\ n \end{pmatrix} \sin^{(m-n)}(\delta) \left(\sum_{j=1}^{(n+1)/2-1} \frac{t^{j+1} f_{n,j}}{(1-2\omega t+t^2)^{2+j}} + \sum_{j=(n+1)/2}^{n} \frac{t^{j+1}\omega'^{2j-n} f_{n,j}}{(1-2\omega t+t^2)^{2+j}} \right),
\end{aligned}$$

where we used the abbreviation ω *instead of* $\omega(\delta,\delta')$ *and* ω' *instead of* $\frac{\partial}{\partial\delta}\omega(\delta,\delta')$, *and* $f_{n,j}$ *are bounded* $\mathcal{C}^\infty$*-functions of* δ *and* δ'.

Proof. The proof is an easy consequence of the following two formulas:

1. n even:

$$\left(\frac{\partial}{\partial\delta}\right)^n \frac{1}{(1-2\omega t+t^2)^2} = \sum_{j=1}^{n/2} \frac{t^j f_{n,j}}{(1-2\omega t+t^2)^{2+j}} + \sum_{j=n/2+1}^{n} \frac{t^j\omega'^{2j-n} f_{n,j}}{(1-2\omega t+t^2)^{2+j}};$$

2. n odd:

$$\left(\frac{\partial}{\partial\delta}\right)^n \frac{1}{(1-2\omega t+t^2)^2} = \sum_{j=1}^{(n+1)/2-1} \frac{t^j f_{n,j}}{(1-2\omega t+t^2)^{2+j}} + \sum_{j=(n+1)/2}^{n} \frac{t^j \omega'^{2j-n} f_{n,j}}{(1-2\omega t+t^2)^{2+j}}.$$

We first look for $n=1$ and $n=2$ and get

$$\begin{aligned}
\frac{\partial}{\partial\delta}\frac{1}{(1-2\omega t+t^2)^2} &= \frac{4t\omega'}{(1-2\omega t+t^2)^{2+1}},\\
f_{1,1}(\delta,\delta') &= 4,\\
\left(\frac{\partial}{\partial\delta}\right)^2 \frac{1}{(1-2\omega t+t^2)^2} &= \frac{4t\omega''}{(1-2\omega t+t^2)^{2+1}} + \frac{24t^2\omega'^2}{(1-2\omega t+t^2)^{2+2}},\\
f_{2,1}(\delta,\delta') &= 4\omega''(\delta,\delta'),\\
f_{2,2}(\delta,\delta') &= 24.
\end{aligned}$$

This shows that our formula is correct for $n=1,2$. We proceed further by induction. We will show only that the formula is correct for $n+1$ if n is even. The other case is totally similar. So we assume that the formula is correct for an even n and get

$$\begin{aligned}
\left(\frac{\partial}{\partial\delta}\right)^{n+1} \frac{1}{(1-2\omega t+t^2)^2} &= \frac{\partial}{\partial\delta}\left(\sum_{j=1}^{n/2} \frac{t^j f_{n,j}}{(1-2\omega t+t^2)^{2+j}} + \sum_{j=n/2+1}^{n} \frac{t^j \omega'^{2j-n} f_{n,j}}{(1-2\omega t+t^2)^{2+j}}\right)\\
&= \sum_{j=1}^{n/2} \frac{t^{j+1}(4+2j)\omega' f_{n,j}}{(1-2\omega t+t^2)^{2+j+1}} + \sum_{j=1}^{n/2} \frac{t^j f'_{n,j}}{(1-2\omega t+t^2)^{2+j}}\\
&\quad + \sum_{j=n/2+1}^{n} \frac{t^{j+1}(4+2j)\omega'^{2j-n+1} f_{n,j}}{(1-2\omega t+t^2)^{2+j+1}}\\
&\quad + \sum_{j=n/2+1}^{n} \frac{t^j (2j-n)\omega'^{2j-n-1}\omega'' f_{n,j}}{(1-2\omega t+t^2)^{2+j}}\\
&\quad + \sum_{j=n/2+1}^{n} \frac{t^j \omega'^{2j-n} f'_{n,j}}{(1-2\omega t+t^2)^{2+j}}\\
&= \sum_{j=2}^{n/2} \frac{t^j (2+2j)\omega' f_{n,j-1}}{(1-2\omega t+t^2)^{2+j}} + \sum_{j=1}^{n/2} \frac{t^j f'_{n,j}}{(1-2\omega t+t^2)^{2+j}}\\
&\quad + \frac{t^{n/2+1=((n+1)+1)/2}(4+2n)\omega' f_{n,n/2}}{(1-2\omega t+t^2)^{2+((n+1)+1)/2}}\\
&\quad + \sum_{j=((n+1)+1)/2+1}^{n+1} \frac{(2+2j)\omega'^{2j-(n+1)} t^j f_{n,j-1}}{(1-2\omega t+t^2)^{2+j}}\\
&\quad + \sum_{j=((n+1)+1)/2}^{n} \frac{t^j \omega'^{2j-(n+1)}(\omega'' f_{n,j} + \omega' f_{n,j})}{(1-2\omega t+t^2)^{2+j}}.
\end{aligned}$$

Because $n/2 = (n+2)/2 - 1$, this proves the formula for $n+1$ if the functions $f_{n+1,j}$ are defined correspondingly. Here we see that $f_{n,j}$ is always a function of the sine and cosine of δ, respectively, δ'. □

The next lemma concerns the derivatives with respect to t.

LEMMA 3.8. *For $n \in \mathbb{N}_0$ we have*

$$\left(\frac{\partial}{\partial t}\right)^{2n} \frac{1}{(1-2\omega t+t^2)^{2+j}} = \sum_{l=0}^{n} a_l^{(j,2n)} \frac{(t-\omega)^{2l}}{((t-\omega)^2+(1-\omega^2))^{2+j+n+l}},$$

$$\left(\frac{\partial}{\partial t}\right)^{2n+1} \frac{1}{(1-2\omega t+t^2)^{2+j}} = \sum_{l=0}^{n} a_l^{(j,2n+1)} \frac{(t-\omega)^{2l+1}}{((t-\omega)^2+(1-\omega^2))^{2+j+n+l+1}}$$

with $a_l^{(j,n)} \in \mathbb{R}$.

An easy consequence is the following formula, which we need in the next lemma:

$$\left(t\frac{\partial}{\partial t}\right)^{k} \left(\frac{t^{j+1}}{(1-2\omega t+t^2)^{2+j}}\right) = \sum_{l=0}^{k} \sum_{i=0}^{\min\{j+1,l\}} b_{j,l,i} t^{j+1+l-i}$$

$$\times \begin{cases} \displaystyle\sum_{p=0}^{(l-i)/2} a_p^{(j,l-i)} \frac{(t-\omega)^{2p}}{((t-\omega)^2+(1-\omega^2))^{2+j+(l-i)/2+p}}, & l-i \textit{ even}, \\ \displaystyle\sum_{p=0}^{(l-i-1)/2} a_p^{(j,l-i)} \frac{(t-\omega)^{2p+1}}{((t-\omega)^2+(1-\omega^2))^{2+j+(l-i+1)/2+p}}, & l-i \textit{ odd}. \end{cases}$$

Proof. We first calculate the first two derivatives and get

$$\frac{\partial}{\partial t} \frac{1}{(1-2\omega t+t^2)^{2+j}} = \frac{-(2+j)2(t-\omega)}{((t-\omega)^2+(1-\omega^2))^{3+j}},$$

$$\left(\frac{\partial}{\partial t}\right)^{2} \frac{1}{(1-2\omega t+t^2)^{2+j}} = -\frac{(4+2j)}{((t-\omega)^2+(1-\omega^2))^{3+j}} + \frac{(4+2j)(6+2j)(t-\omega)^2}{((t-\omega)^2+(1-\omega^2))^{4+j}},$$

which proves our result with the corresponding definitions of $a_0^{(j,1)}, a_0^{(j,2)}$, and $a_1^{(j,2)}$. We proceed further by induction, but we again prove only one case. We assume that the formula is correct for $2n+1$ and prove it for $2(n+1)$. By our assumption we have

$$\begin{aligned} \left(\frac{\partial}{\partial t}\right)^{2(n+1)} \frac{1}{(1-2\omega t+t^2)^{2+j}} &= \sum_{l=0}^{n} \Bigg(\frac{a_l^{(j,2n+1)}(2l+1)(t-\omega)^{2l}}{((t-\omega)^2+(1-\omega^2))^{2+j+n+l+1}} \\ &\qquad - \frac{a_l^{(j,2n+1)} 2(t-\omega)^{2l+2}(2+j+n+1+l)}{((t-\omega)^2+(1-\omega^2))^{2+j+(n+1)+(l+1)}} \Bigg) \\ &= \sum_{l=0}^{n} \frac{a_l^{(j,2n+1)}(2l+1)(t-\omega)^{2l}}{((t-\omega)^2+(1-\omega^2))^{2+j+(n+1)+l}} \\ &\qquad - \sum_{l=1}^{n+1} \frac{a_{l-1}^{(j,2n+1)} 2(t-\omega)^{2l}(2+j+n+l)}{((t-\omega)^2+(1-\omega^2))^{2+j+(n+1)+l}}, \end{aligned}$$

and this proves our formula for $2(n+1)$ if the coefficients are defined correctly. For the proof of the last formula we use

$$\left(t\frac{\partial}{\partial t}\right)^{j} g(t) = \sum_{i=1}^{j} c_{j,i} t^i \left(\frac{\partial}{\partial t}\right)^{i} g(t), \quad c_{j,i} \in \mathbb{R},$$

and the formula

$$\left(\frac{\partial}{\partial t}\right)^j (fg)(t) = \sum_{i=0}^{j} \left(\begin{array}{c} j \\ i \end{array} \right) f^{(i)}(t) g^{(j-1)}(t). \qquad \square$$

Now we are ready to prove our main result on the kernel $\overline{k}_1$.

LEMMA 3.9. *For $m, k \in \mathbb{N}_0$ we have*

$$\sin(\delta)^{m+k} \left(t\frac{\partial}{\partial t}\right)^k \left(\frac{\partial}{\partial \delta}\right)^m \overline{k}_1(\delta, \delta', t) = g_{m,k}(\delta, \delta', t)\overline{k}_1(\delta, \delta', t)$$

with a bounded continuous function $g_{m,k}$ on $[0, \delta_0]^2 \times [0, \infty)$.

Proof. We have only to consider $\overline{k}_2$. By the previous two lemmas we get

$$\begin{aligned}
&\sin(\delta)^{m+k} \left(t\frac{\partial}{\partial t}\right)^k \left(\frac{\partial}{\partial \delta}\right)^m \overline{k}_2(\delta, \delta', t) \\
&= \sum_{n=0, 2|n}^{m} \left(\begin{array}{c} m \\ n \end{array} \right) \sum_{j=1}^{n/2} f_{n,j} \sum_{l=1}^{k} \sum_{i=0}^{\min\{j+1,l\}} b_{j,l,i}\alpha_{m,n,j,l,i} \\
&\quad + \sum_{n=0, 2|n}^{m} \left(\begin{array}{c} m \\ n \end{array} \right) \sum_{j=n/2+1}^{n} f_{n,j} \sum_{l=1}^{k} \sum_{i=0}^{\min\{j+1,l\}} b_{j,l,i}\omega'^{2j-n}\alpha_{m,n,j,l,i} \\
&= \sum_{n=0, 2\nmid n}^{m} \left(\begin{array}{c} m \\ n \end{array} \right) \sum_{j=1}^{(n+1)/2-1} f_{n,j} \sum_{l=1}^{k} \sum_{i=0}^{\min\{j+1,l\}} b_{j,l,i}\alpha_{m,n,j,l,i} \\
&\quad + \sum_{n=0, 2\nmid n}^{m} \left(\begin{array}{c} m \\ n \end{array} \right) \sum_{j=n/2+1}^{n} f_{n,j} \sum_{l=1}^{k} \sum_{i=0}^{\min\{j+1,l\}} b_{j,l,i}\omega'^{2j-n}\alpha_{m,n,j,l,i},
\end{aligned}$$

where

$$\begin{aligned}
\alpha_{m,n,j,l,i} &= \sin(\delta)^{m+k} \sin^{(m-n)}(\delta) t^{j+1+l-i} \\
&\quad \times \begin{cases} \displaystyle\sum_{p=0}^{(l-i)/2} \frac{a_p^{(j,l-i)}(t-\omega)^{2p}}{((t-\omega)^2 + (1-\omega^2))^{2+j+(l-i)/2+p}}, & l-i \text{ even}, \\ \displaystyle\sum_{p=0}^{(l-i-1)/2} \frac{a_p^{(j,l-i)}(t-\omega)^{2p+1}}{((t-\omega)^2 + (1-\omega^2))^{2+j+(l-i+1)/2+p}}, & l-i \text{ odd}. \end{cases}
\end{aligned}$$

Now one has to consider 16 different summands ($\alpha_{m,n,j,l,i}$ for m even/odd, $l-i$ even/odd). We will consider only the two cases m even, n even, $j \geq n/2+1$, and $l-i$ even and odd. This corresponds to the two cases in the second of the above sums. We have only to prove our lemma for small t because for large t the formula is clearly correct.

First, for $l-i$ even and p arbitrary,

$$\begin{aligned}
\alpha_{m,n,j,l,i}(\delta, \delta', t) &= \frac{\sin(\delta)^{m+k}(\pm\sin(\delta))t^{j+1+l-i}(t-\omega)^{2p}\omega'^{2j-n}}{((t-\omega)^2 + (1-\omega^2))^{2+j+(l-i)/2+p}} \\
&= \pm\overline{k}_2 \frac{\sin(\delta)^{m+k}t^{j+l-i}\omega'^{2j-n}}{((t-\omega)^2 + (1-\omega^2))^{j+(l-i)/2}} \frac{(t-\omega)^{2p}}{((t-\omega)^2 + (1-\omega^2))^p} \\
&\leq C_1\overline{k}_2 \frac{\sin(\delta)^{m+k}t^{j+l-i}\omega'^{2j-n}}{((t-\omega)^2 + (1-\omega^2))^{j+(l-i)/2}}
\end{aligned}$$

with some constant $C_1 > 0$. Now we remind the reader of

$$\begin{aligned} \omega'(\delta,\delta') &\leq C_2(\sin(\delta)+\sin(\delta')), && (3.38)\\ (1-2\omega t+t^2) &= (t-1)^2+2t(1-\omega) \\ &\geq 2t(1-\omega)\ \ldots \\ &\geq C_3 t(\sin(\delta)+\sin(\delta'))^2, && (3.39) \end{aligned}$$

where we have again used the addition theorems for sine and cosine in the last formula. We further estimate

$$\frac{\sin(\delta)^{m+k}t^{j+l-i}\omega'^{2j-n}}{((t-\omega)^2+(1-\omega^2))^{j+(l-i)/2}} \overset{(3.38),(3.39)}{\leq} C_4\frac{t^{(l-i)/2}\sin(\delta)^{m+k}}{(\sin(\delta)+\sin(\delta'))^{n+l-i}},$$

$C_4 > 0$, and for the exponent of the denominator we get

$$\begin{aligned} n+l-i &\leq n+l \\ &\leq m+k. \end{aligned}$$

This proves the result.

Second, for $l-i$ odd and p arbitrary,

$$\begin{aligned} \alpha_{m,n,j,l,i}(\delta,\delta',t) &= \frac{\sin(\delta)^{m+k}(\pm\sin(\delta))t^{j+1+l-i}(t-\omega)^{2p+1}\omega'^{2j-n}}{((t-\omega)^2+(1-\omega^2))^{2+j+(l-i+1)/2+p}} \\ &= \pm\bar{k}_2\frac{\sin(\delta)^{m+k}t^{j+l-i}\omega'^{2j-n}}{((t-\omega)^2+(1-\omega^2))^{j+(l-i)/2}}\frac{(t-\omega)^{2p+1}}{((t-\omega)^2+(1-\omega^2))^{p+1/2}} \\ &\leq C_5\bar{k}_2\frac{\sin(\delta)^{m+k}t^{j+l-i}\omega'^{2j-n}}{((t-\omega)^2+(1-\omega^2))^{j+(l-i)/2}} \end{aligned}$$

with some constant $C_5 > 0$. Now we are in the same situation as above, and our proof is finished. □

We will now introduce the function spaces $X^m_{\nu_1,\nu_2}$ for each face F_j of the cone Γ. After a suitable rotation we can assume that F_j has the following representation:

$$F_j = \left\{ r\begin{pmatrix} \sin(\delta) \\ 0 \\ \cos(\delta) \end{pmatrix} \mid 0 \leq r < \infty,\ \delta\in[0,\delta_0] \right\}, \quad \delta_0 \in (0,2\pi).$$

Then the norm on the function space $X^m_{\nu_1,\nu_2}(F_j)$ is defined by

$$(3.40)\quad \|u; X^m_{\nu_1,\nu_2}\|^2$$
$$:= \sum_{i+l\leq m}\int_0^\infty\int_0^{\delta_0} r^{-2\nu_1}\sin\left(\delta\frac{\pi}{\delta_0}\right)^{-2\nu_2}\left(\sin\left(\delta\frac{\pi}{\delta_0}\right)^{i+l}\left(r\frac{\partial}{\partial r}\right)^i\left(\frac{\partial}{\partial\delta}\right)^l u\right)^2 d\delta\, dr.$$

The above lemma implies the following corollary.

Corollary 3.10. *Let F_j and F_{j+1} be two adjacent faces of the infinite cone, and define L to be the integral operator*

$$Lu := \widetilde{K}u|_{F_{j+1}};$$

here the function u is given on F_j and extended by zero to the whole of Γ, *and* $\widetilde{K}$ *is the radiosity operator on the infinite cone; see* (2.21). *Then we get*

$$L:\, X^0_{\nu_1,\nu_2}(F_j) \longrightarrow X^m_{\nu_1,\nu_2}(F_{j+1})$$

for all $m \in \mathbb{N}$, *and* $(\nu_1, \nu_2) \in (-2.5, 1.5) \times (-1.5, 1.5)$.

Proof. The proof follows from the previous lemma and Theorem 3.6. □

To formulate a further application of Lemma 3.9, we introduce a weighted Sobolev space $X^m_{\nu_1,\nu_2}(S)$, $m \in \mathbb{N}$, on S in the following way. Every face Δ_j of S has three vertices, which we will denote by $v^{(j)}_i$, $i = 1, 2, 3$. There exists a linear transformation $\mathcal{T}_{j,i}$ (i.e., a rotation followed by a translation) with

$$\mathcal{T}_{j,i}(v^{(j)}_i) = 0,$$

and the infinite triangle $F_{j,i}$ generated by $\mathcal{T}_{j,i}\Delta_j$ fulfills

$$F_{j,i} = \left\{ r \begin{pmatrix} \sin(\delta) \\ 0 \\ \cos(\delta) \end{pmatrix} \mid 0 \le r < \infty,\, \delta \in [0, \delta_{j,i}] \right\}, \quad \delta_{j,i} \in (0, 2\pi). \tag{3.41}$$

For each S_j there exist three $\mathcal{C}^\infty$-functions $\varphi_{j,i}$, $i = 1, 2, 3$, which are nonnegative, and

$$\begin{aligned} \varphi_{j,1}(x) + \varphi_{j,2}(x) + \varphi_{j,3}(x) = 1, &\quad x \in \Delta_j, \\ \varphi_{j,i}|_{U_\varepsilon(v^{(j)}_i)} \equiv 1, &\quad \varepsilon > 0. \end{aligned}$$

Further, the support of $\varphi_{j,i}$ does not intersect the edge of Δ_j on the opposite side of $v^{(j)}_i$. Now $X^m_{\nu_1,\nu_2}(S)$ is defined by the norm

$$\|u :\, X^m_{\nu_1,\nu_2}(S)\|^2 := \sum_{j=1}^{n} \sum_{i=1}^{3} \|(\varphi_{j,i} u) \circ \mathcal{T}^{-1}_{j,i}; X^m_{\nu_1,\nu_2}(F_{j,i})\|^2. \tag{3.42}$$

Now we can formulate a special result which follows easily from Lemma 3.9 and shows its implications if the shadow lines do not disturb the regularity.

COROLLARY 3.11. *Let S be convex, $E \in L^2(S)$ with $E|_{\Delta_j} \in \mathcal{C}^\infty(\Delta_j)$, $j = 1(1)n$; then the solution $u \in L^2(S)$ of* (1.1) *fulfills*

$$u \in X^m_{\nu_1,\nu_2}(S), \quad m \in \mathbb{N}, \tag{3.43}$$

$\nu_1 \in (-2.5, -0.5]$, $\nu_2 \in (-1.5, 0]$.

Proof. We remark that the solution $u \in L^2(S)$ belongs to $X^0_{\nu_1,\nu_2}(S)$, (ν_1, ν_2) as above and $f \in X^m_{\nu_1,\nu_2}(S)$, (ν_1, ν_2) as above. The representation

$$u = f - Ku$$

for u and Lemma 3.9 prove the result. □

Remark. We mention here that it would be more interesting to get similar results for $\nu_1, \nu_2 > 0$. This seems to be possible. Because of Theorem 3.6, one gets the invertibility of the local operators for positive ν values, and by some localization techniques this shows that the operator $(I - K)$ is a Fredholm operator with index zero in $X^0_{\nu_1,\nu_2}(S)$, $\nu_1, \nu_2 > 0$, small enough. But because of $X^0_{\nu_1,\nu_2} \subset L^2(S)$ and kernel$(I-K)|_{L^2(S)} = \{0\}$ one gets the invertibility. This would show Corollary 3.11 also for some positive values of ν_1, ν_2.

However, Lemma 3.9 also makes it possible to get pointwise estimates for the solution of (1.1) if the assumptions of the above corollary are fulfilled; see [18, Lemma 5.2].[1]

COROLLARY 3.12. *Let S be convex, $E \in L^2(S)$, $E|_{\Delta_j} \in \mathcal{C}^\infty(\Delta_j)$, $j = 1(1)n$, and $u \in L^2(S)$ be the solution of* (1.1). *Let Δ_j, $j \in \{1,\dots,n\}$, be an arbitrary face with vertex $v_i^{(j)}$, $i \in \{1,2,3\}$, $D_{j,i} := [0,\infty) \times [0, \delta_{j,i}/2]$ (see* (3.41)*), and define*

$$u_{j,i} \;:=\; (\varphi_{j,i} u) \circ \mathcal{T}_{j,i}^{-1} \;:\; D_{j,i} \longrightarrow \mathbb{R}.$$

Then we get

$$\left.\begin{array}{lll} \sup_{(r,\delta)\in D_{j,i}} \left| \left(\sin(\delta)\frac{\partial}{\partial\delta}\right)^l u_{j,i}(r,\delta)\right| & \leq & c_l \\ \sup_{(r,\delta)\in D_{j,i}} \left| \left(r\frac{\partial}{\partial r}\right)^l u_{j,i}(r,\delta)\right| & \leq & c_l \end{array}\right\}, \qquad l \in \mathbb{N}_0,\ c_l > 0. \tag{3.44}$$

Proof. The proof is analogous to the proof of Rathsfeld's lemma. First, we notice that in the case of a convex domain the right-hand side of (2.24), and also of (3.8), is a $\mathcal{C}^\infty$-function. So we have to prove that

$$u_{j,i} := A^{-1}y,\ A \,:=\, I - \overline{K},$$

y a $\mathcal{C}^\infty$-function, fulfills (3.44). But the operator $\overline{K}$ fulfills the estimate (2.9) of Lemma 2.1. This also implies $\|\overline{K}\|_{L^\infty(D_{j,i})} < 1$; see [16]. But then we get

$$A^{-1} \;=\; \sum_{j=0}^{\infty} \overline{K}^j \;:\; L^\infty(D_{j,i}) \longrightarrow L^\infty(D_{j,i})$$

and

$$A^{-1} = I + \overline{K}A^{-1}.$$

The operator $\overline{K}$ is a Mellin convolution operator with respect to r, and so

$$\left(r\frac{\partial}{\partial r}\right)\overline{K}u = \overline{K}\left(r\frac{\partial}{\partial r}u\right).$$

Now

$$\left(r\frac{\partial}{\partial r}\right)^l u_{j,i} = A^{-1}\left(r\frac{\partial}{\partial r}y\right)$$

is bounded and

$$\left(\sin(\delta)\frac{\partial}{\partial\delta}\right)^l u_{j,i} = \left(\sin(\delta)\frac{\partial}{\partial\delta}\right)^l y + \underbrace{\left[\left(\sin(\delta)\frac{\partial}{\partial\delta}\right)^l \overline{K}\right]}_{=:\overline{K}_l} A^{-1}y.$$

The kernel of the $\overline{K}_l$ has the same structure as the kernel $\overline{k}_1$ of $\overline{K}$ by Lemma 3.9, so it maps $L^\infty(D_{j,i})$ into $L^\infty(D_{j,i})$, and we have proved our corollary. ∎

[1] The author would like to thank Dr. J. Elschner for the reference to this lemma.

4. The regularity near the vertices. Now we start to study the regularity of the right-hand sides of (3.3) and (2.23). The difference between these two equations is that the right-hand sides of (3.3) also contain some contributions from the local cone which come from the nonadjacent faces. But first we consider the function f_2 in (2.23).

For each triangle Δ_j, $j \in \{1, \ldots, N\}$ (see (2.1)) we denote by $\{v_1^{(j)}, v_2^{(j)}, v_3^{(j)}\}$ the vertices and by $\{e_1^{(j)}, e_2^{(j)}, e_3^{(j)}\}$ the edges of Δ_j. Given a triangle $\Delta \subset \mathbb{R}^3$, we denote by $E(\Delta)$ the plane which is spanned by Δ. If $x \in \mathbb{R}^3$ and $l \subset \mathbb{R}^3$ is a segment in $\mathbb{R}^3$, we denote

$$E(x,l) := \begin{cases} \text{the plane spanned by } l \text{ and } x & \text{if } x \text{ is not an element of the line} \\ & \text{spanned by } l, \\ \text{the line spanned by } l & \text{otherwise.} \end{cases}$$

For $j \in \{1, \ldots, N\}$ we denote by T_j the subset of S where the shadow lines created by Δ_j could cause problems for the regularity of the solution of the radiosity equation. To define T_j we need some preparations. Let

$$D_{j,m} := \begin{cases} E(\Delta_j) \cap \Delta_m & \text{if some parts of } \Delta_j \text{ are visible from } \Delta_m, \\ \emptyset & \text{otherwise,} \end{cases}$$

$m \in \{1, \ldots, N\} \setminus \{j\}$, and

$$\widetilde{D}_{j,m,k,l,i} := \begin{cases} E(e_k^{(j)}, v_l^{(i)}) \cap \Delta_m & \text{if } v_k^{(j)} \text{ or some parts of } e_l^{(i)} \text{ can be seen from } \Delta_m, \\ \emptyset & \text{otherwise,} \end{cases}$$

$k, l \in \{1, 2, 3\}$, $i \in \{1, \ldots, N\} \setminus \{j\}$. From a point $x \in D_{j,m} \subset \Delta_m$ one "sees" the set Δ_j only as a line. Assume $x \in D_{j,m,k,l,i} \subset \Delta_m$; if one "stands" at the point x and looks in the direction of the l vertex of Δ_i, the ray to the vertex also hits the edge number k of side Δ_j. So it is clear that in every neighborhood of x one finds points which see the vertex $v_l^{(i)}$ and others which do not see it. This indicates that the shape of the shadow (it is a polygon), seen from x, changes essentially (the number of vertices of the polygon) in the vicinity of x. In the above definitions the phrase "some parts of set A can be seen from Δ_m" means that $\vec{n}_A \cdot \vec{n}_{\Delta_m} < 0$, and there exists a point $y \in A$ and $x \in \Delta_m$ with $(y - x) \cdot \vec{n}_{\Delta_m} > 0$. Here the choice of the point x plays no role. Now T_j is given by

$$T_j := \bigcup_{m=1,\, m \neq j}^{N} \left(D_{j,m} \cup \left(\bigcup_{k,l=1}^{3} \bigcup_{i=1, i \notin \{m,j\}}^{N} \widetilde{D}_{j,m,k,l,i} \right) \right). \tag{4.1}$$

T_j is the set of points on S, where the shadows of Δ_j can cause problems, and T is the union of all these critical lines

$$T := \bigcup_{j=1}^{N} T_j. \tag{4.2}$$

Because T is the union of lines, we get

$$S' := \dot{S} \setminus T = \bigcup_{j=1}^{N'} \Delta'_j, \tag{4.3}$$

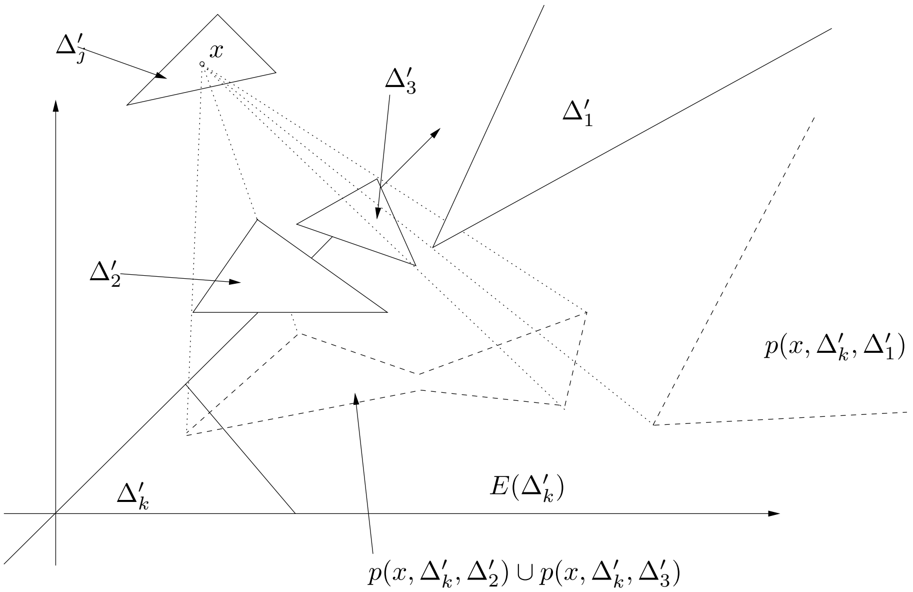

Δ'_j
x
Δ'_3
Δ'_1
Δ'_2
$p(x,\Delta'_k,\Delta'_1)$
Δ'_k
$E(\Delta'_k)$
$p(x,\Delta'_k,\Delta'_2)\cup p(x,\Delta'_k,\Delta'_3)$

FIG. 4. *The construction for projections* $p(x,\Delta'_k,\cdot)$.

where the triangles Δ'_j are open.

To explain the meaning of T we now fix $j,k\in\{1,\dots,N'\}$, $j\neq k$, and consider the shadows of S' seen from Δ'_j on $E(\Delta'_k)$. For $x\in\Delta'_j$ we denote by $p(x,\Delta'_k,\Delta'_i)$ the projection of the points $y\in\Delta'_i$ on $E(\Delta'_k)$. This means that if $\vec{n}_x\cdot\vec{n}_y<0$ and $(y-x)\cdot\vec{n}_x>0$ and if there is a $\lambda\geq 1$ such that $x+\lambda(y-x)\in E(\Delta'_k)$, then $x+\lambda(y-x)\in p(x,\Delta'_k,\Delta'_i)$. For each i, $p(x,\Delta'_k,\Delta'_i)$ is either empty, a triangle, or a triangle where one or two corners are in infinity. But it is never a line because of the definition of T. Now for $p(x,\Delta'_k,\Delta'_i)\neq\emptyset$ and $p(x,\Delta'_k,\Delta'_l)\neq\emptyset$, $i\neq l$, no vertex of $p(x,\Delta'_k,\Delta'_i)$ crosses the boundary of $p(x,\Delta' k,\Delta'_l)$ as x varies in Δ'_j. Also, no edge of $p(x,\Delta'_k,\Delta_i)$ crosses parallel to an edge of $p(x,\Delta' k,\Delta'_l)$. So $p(x,\Delta'_k,\Delta'_i)\cup p(x,\Delta'_k,\Delta'_l)$ are either two triangles or one polygonal domain in $E(\Delta'_k)$. This structure stays stable for all $x\in\Delta'_j$. The vertices of this triangle or polygonal domain are $\mathcal{C}^\infty$-functions on Δ'_j.

Now we can add in succession all shadows and get

$$\begin{aligned} P_{j,k} &:= \widetilde{\bigcup}_{i=1,i\notin\{j,k\}}^{N'} p(x,\Delta'_k,\Delta'_i) \\ &= \bigcup_{i=1}^{N'(j,k)} \widetilde{p}_{j,k,i}(x), \end{aligned} \tag{4.4}$$

where the union in the first line does not contain the Δ' which are adjacent to Δ'_j. Each $\widetilde{p}_{j,k,i}(x)$ is a polygonal domain in $E(\Delta'_k)$, the edges of $\widetilde{p}_{j,k,i}(x)$ are $\mathcal{C}^\infty$-functions on Δ'_j, and $\widetilde{p}_{j,k,i}(x)\cap\widetilde{p}_{j,k,i'}(x)=\emptyset$, $i\neq i'$. If we denote by $\widetilde{v}_{j,k,i}(x)$ an arbitrary vertex of $\widetilde{p}_{j,k,i}(x)$, we get for $x,y\in\Delta'_j$ that the curve

$$\lambda\longrightarrow\widetilde{v}_{j,k,i}(x+\lambda(y-x)),\quad \lambda\in[0,1],$$

moves along a straight line, and the velocity is always greater than zero (see Figure 4).

Lemma 4.1. *Let $u : S \to \mathbb{R}$, $u|_{\Delta'_j}$ be continuous, $j = 1(1)N'$. Let $j_0 \in \{1, \dots, N'\}$, and define S_{j_0} to be the union of all Δ'_j which are not equal or adjacent to Δ'_{j_0}. Define*

$$w(x) := \int_{S_{j_0}} \beta(x,y) \frac{\vec{n}_x \cdot (y-x)\, \vec{n}_y \cdot (x-y)}{\|x-y\|^4} u(y)dy;$$

then we get $w \in \mathcal{C}^1(\overline{\Delta'_{j_0}})$.

Proof. It is sufficient to prove that

$$w(x) = \int_{\Delta'_j} \beta(x,y) \frac{\vec{n}_x \cdot (y-x)\, \vec{n}_y \cdot (x-y)}{\|x-y\|^4} u(y)dy$$

is in $\mathcal{C}^1(\overline{\Delta'_{j_0}})$, where Δ'_j is one of the admissible triangles; therefore,

$$k(x,y) := \frac{\vec{n}_x \cdot (y-x)\, \vec{n}_y \cdot (x-y)}{\|x-y\|^4}$$

is a $\mathcal{C}^\infty$-function on $\Delta'_{j_0} \times \Delta'_j$. Define

$$\Delta'_j(x) := \Delta'_j \setminus \left(\bigcup_{k=1}^{N(j_0,j)} \widetilde{p}_{j_0,j,k}(x) \right).$$

Then

$$w(x) = \int_{\Delta'_j(x)} \frac{\vec{n}_x \cdot (y-x)\, \vec{n}_y \cdot (x-y)}{\|x-y\|^4} u(y)dy.$$

Because the corners of $\Delta'_j(x)$ are $\mathcal{C}^\infty$-functions of x, we find for every x a real $\delta > 0$ and a triangulation

$$\Delta'_j = \bigcup_{i=1}^{M_j} \Delta''_i$$

(see Figure 5), such that for all $z \in U_\delta(x) \cap \Delta'_{j_0}$ in every Δ''_i there is at most one corner point of $P_{j_0,j}$, and this point does not move out of Δ''_i when z varies in $U_\delta(x)$.

This implies that $\Delta''_i(z) := \Delta''_i \cap \Delta'_j(z)$ is either empty or has one of the following forms (see Figure 6).

If $\Delta''_i(z)$ is empty or has the form of case (a) for all $z \in U_\delta(x)$, it follows that

$$\int_{\Delta''_i(z)} k(z,y)u(y)dy$$

is a $\mathcal{C}^\infty$-function of z. The cases (b) and (d) and the cases (c) and (e) are complementary, and we have only to look for the cases (b) and (c). After some smooth transformation we can assume that we have the situation shown in Figure 7.

Here ξ_0, ξ_1, η_1, and η_2 are $\mathcal{C}^\infty$-functions of $z \in U_\delta(x)$. We will treat only the left case and give the result for the right case. We further consider z to be a real (one dimensional) variable because no direction has a special significance, and we will see that the first derivative is continuous.

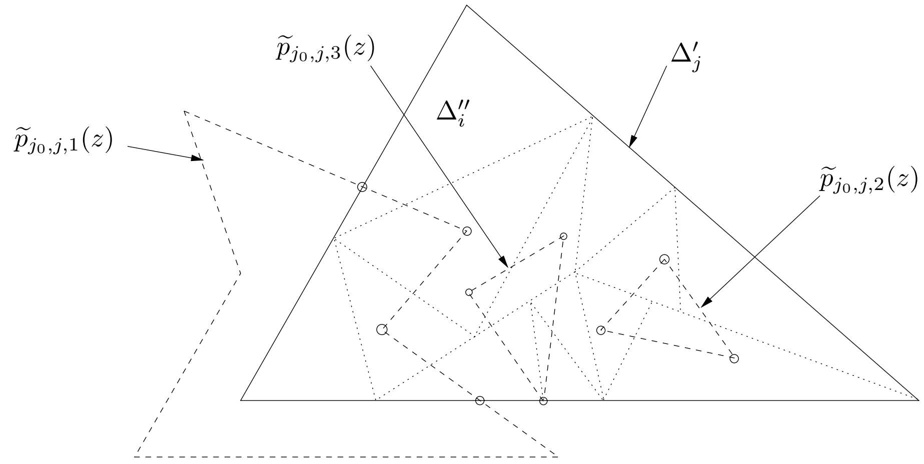
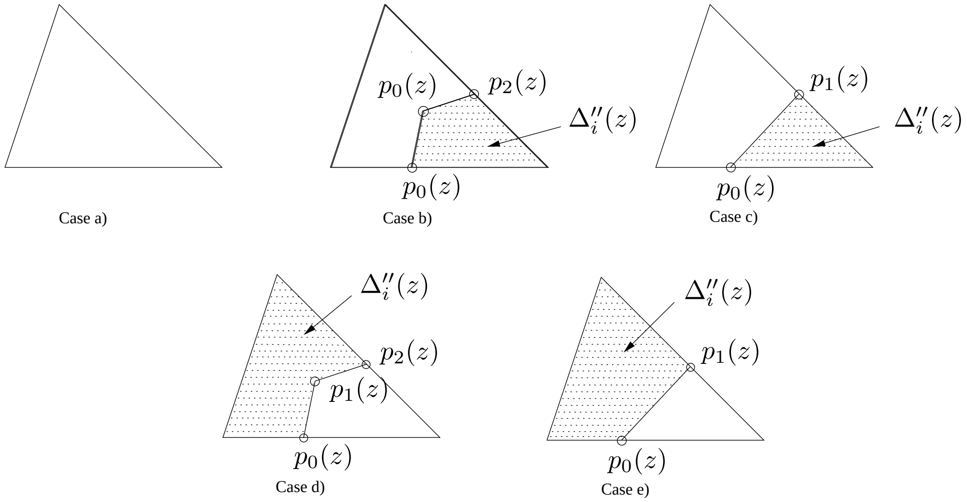

$\widetilde{p}_{j_0,j,3}(z)$
Δ'_j
Δ''_i
$\widetilde{p}_{j_0,j,1}(z)$
$\widetilde{p}_{j_0,j,2}(z)$

FIG. 5. *The partition of* $\Delta'_j(z)$.



Case a)

Case b)

Case c)

Case d)

Case e)

FIG. 6. *The possible shapes for* $\Delta''_i(z)$.

Define

$$\begin{aligned} F(z) &:= \int_{\Delta''_i(z)} k(z,\xi,\eta)u(\xi,\eta)d\xi d\eta \\ &= \int_0^{\xi_1(z)} \int_0^{s_1(z,\xi)} k(z,\xi,\eta)u(\xi,\eta)d\eta d\xi + \int_0^{\eta_1(z)} \int_{\eta_1(z)}^{s_2(z,\eta)} k(z,\xi,\eta)u(\xi,\eta)d\xi d\eta, \end{aligned}$$

where

$$\begin{aligned} s_1(z,\xi) &:= \eta_2(z) + \xi\frac{\eta_1(z)-\eta_2(z)}{\xi_1(z)}, \\ s'_1(z,\xi) &:= \frac{\partial}{\partial z} s_1(z,\xi) \end{aligned}$$

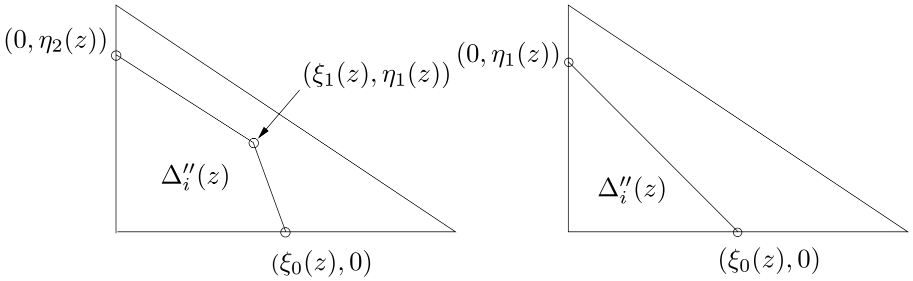


FIG. 7. *The two remaining possible shapes for* $\Delta_i''(z)$.

$$= \eta_2'(z) + \xi \frac{(\eta_1'(z)-\eta_2'(z))\xi_1(z) - \xi_1'(z)(\eta_1(z)-\eta_2(z))}{\xi_1^2(z)},$$

and

$$\begin{aligned} s_2(z,\eta) &:= \xi_0(z) + \eta \frac{\xi_1(z)-\xi_0(z)}{\eta_1(z)}, \\ s_2'(z,\eta) &= \frac{\partial}{\partial z} s_2(z,\eta) \\ &= \xi_0'(z) + \eta \frac{(\xi_1'(z)-\xi_0'(z))\eta_1(z) - \eta_1'(z)(\xi_1(z)-\xi_0(z))}{\eta_1^2(z)}. \end{aligned}$$

Now

$$\begin{aligned} F(z+h) - F(z) = &\int_0^{\xi_1(z)} \int_{s_1(z,\xi)}^{s_1(z+h,\xi)} k(z+h,\xi,\eta)u(\xi,\eta)d\eta d\xi \\ &+ \int_0^{\xi_1(z)} \int_0^{s_1(z,\xi)} (k(z+h,\xi,\eta) - k(z,\xi,\eta))u(\xi,\eta)d\eta d\xi \\ &+ \int_0^{\eta_1(z)} \int_{s_2(z,\eta)}^{s_2(z+h,\eta)} k(z+h,\xi,\eta)u(\xi,\eta)d\xi d\eta \\ &+ \int_0^{\eta_1(z)} \int_{\eta_1(z)}^{s_2(z,\eta)} (k(z+h,\xi,\eta) - k(z,\xi,\eta))u(\xi,\eta)d\xi d\eta \;+\; O(h^2). \end{aligned}$$

This implies

$$\begin{aligned} F'(z) = &\int_0^{\xi_1(z)} k(z,\xi,s_1(z,\xi))u(\xi,s_1(z,\xi))s_1'(z,\xi)d\xi \\ &+ \int_0^{\eta_1(z)} k(z,s_2(z,\eta))u(s_2(z,\eta),\eta)s_2'(z,\eta)d\eta \\ &+ \int_{\Delta_i''(z)} \frac{\partial k}{\partial z}(z,\xi,\eta)u(\xi,\eta)d\xi d\eta, \end{aligned}$$

and this function depends continuously on the variable z. But because of the appearance of $u(\xi,s_1(z,\xi))$ and $u(s_2(z,\eta),\eta)$ under the integral sign, we cannot expect that this function is differentiable if u is only continuous.

In the case of the right-hand side of the picture, we get

$$F'(z) = \int_0^{\xi_0(z)} k(z,\xi,s_1(z,\xi))u(\xi,s_1(z,\xi))s_1'(z,\xi)d\xi$$

$$+\int_{\Delta_i''(z)} \frac{\partial k}{\partial z}(z,\xi,\eta)u(\xi,\eta)d\eta d\xi,$$

and this shows the result for the second case. □

Remark. The previous lemma characterizes the smoothness of the function f_2 in (2.23) because the support of $(1-\varphi)\overline{u}$ is separated from the support of ψ, and we can apply Lemma 4.1 to derive the behavior of f_2. For a more thorough discussion and data structures for the calculation of the triangulation $\cup_{j=1}^{N'}\Delta_j'$, see [6].

Now we turn to f_3 in (2.23). For $x \in B_{2/3}(0)$ the function f_3 is given by

$$f_3(x) = \widetilde{K}[(1-\psi)\varphi u](x),$$

and the smoothness of f_3 follows from the above discussion. For $|x| \geq 2/3$ the smoothness of f_3 follows from the discussion of Rathsfeld [17, Lemma 4.1(ii)], where he showed that this term is in $\mathcal{C}^1$ for continuous u.

For the following we assume that either $T \cap B_1(0) = \emptyset$ (see (4.2)) or $0 \in T$, and then we assume that we introduce new faces on Γ (see (2.20)) to guarantee that each line of $T \cap \Gamma$ belongs to an edge. The last contribution to the right-hand side in (3.3) are the integrals of the faces of Γ, which are not adjacent, or the corresponding parts of the spherical γ are separated. Let F_1 and F_2 be two nonadjacent faces of Γ. We again assume that F_1 is given by

$$(4.5)\quad F_1 := \left\{ r\begin{pmatrix} \sin(\delta) \\ 0 \\ \cos(\delta) \end{pmatrix} =: p_1(r,\delta) \,|\, 0 \leq r \leq \infty,\ \delta \in [0,\delta_0] \right\} \text{with normal}\, \vec{n}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

For F_2 we assume that the normal is given by

$$(4.6)\qquad \vec{n}_2 := \begin{pmatrix} \sin(\alpha)\cos(\vartheta) \\ \sin(\alpha)\sin(\vartheta) \\ \cos(\alpha) \end{pmatrix}, \quad \alpha \in [0,\pi),\ \vartheta \in [-\pi,\pi).$$

To satisfy the visibility assumption, ϑ has to fulfill $\vartheta \in (-\pi, 0)$. For the plane perpendicular to $\vec{n}_2$ we have the two orthogonal vectors

$$\begin{pmatrix} \cos(\alpha)\cos(\vartheta) \\ \cos(\alpha)\sin(\vartheta) \\ -\sin(\alpha) \end{pmatrix}, \begin{pmatrix} -\sin(\vartheta) \\ \cos(\vartheta) \\ 0 \end{pmatrix}.$$

Therefore, we can assume

$$(4.7)\qquad F_2 = \left\{ r\left[\cos(\delta)\begin{pmatrix} \cos(\alpha)\cos(\vartheta) \\ \cos(\alpha)\sin(\vartheta) \\ -\sin(\alpha) \end{pmatrix} + \sin(\delta)\begin{pmatrix} -\sin(\vartheta) \\ \cos(\vartheta) \\ 0 \end{pmatrix}\right] =: p_2(r,\delta) \,|\, r \geq 0,\ \delta \in [\delta_1,\delta_2] \right\}.$$

Let $u \in \mathcal{C}([0,\infty) \times [\delta_1,\delta_2])$ with compact support, and define

$$(4.8)\ w(r,\delta) := \int_{F_2} \beta(r,\delta,r',\delta') \frac{\vec{n}_1 \cdot (p_2(r',\delta') - p_1(r,\delta))\, \vec{n}_2 \cdot (p_1(r,\delta) - p_2(r',\delta'))}{\|p_1(r,\delta) - p_2(r',\delta')\|^4} r' u(r',\delta') dr' d\delta'$$
$$= \int_0^\infty \int_{\delta_1}^{\delta_2} \beta(r,\delta,r',\delta') k(r,\delta,r',\delta') u(r',\delta') d\delta' dr',$$

where

$$(4.9)\ k(r,\delta,r',\delta') = r'^2 r \frac{(\cos(\delta)\cos(\alpha)\sin(\vartheta) + \sin(\delta)\cos(\vartheta))(\sin(\delta)\sin(\alpha)\cos(\vartheta) + \cos(\delta)\cos(\alpha))}{(r^2 + r'^2 - 2rr'\omega(\delta,\delta'))^2}$$

and

$$\omega(\delta,\delta') := \sin(\delta)(\cos(\delta')\cos(\alpha)\cos(\vartheta) - \sin(\delta')\sin(\vartheta)) - \cos(\delta)\cos(\delta')\sin(\alpha)$$
$$(4.10)\qquad \leq c < 1, \quad \delta \in [0,\delta_0], \quad \delta' \in [\delta_1,\delta_2],$$

because F_1 and F_2 are not adjacent. This implies

$$(4.11)\qquad k \in \mathcal{C}^\infty([0,\infty) \times [0,\delta_0] \times [0,\infty) \times [\delta_1,\delta_2]).$$

If we take into account shadow lines, then we get

$$(4.12)\qquad w(r,\delta) = \int_0^\infty \int_{\delta_1(r,\delta)}^{\delta_2(r,\delta)} k(r,\delta,r',\delta') u(r',\delta') d\delta' dr'.$$

This shows that

$$(4.13)\qquad w \in \mathcal{C}^1([0,\infty) \times [0,\delta]),$$

where we mention also that the first derivative is continuous up to the boundary.

If the right-hand side $E(x)$ of (1.1) is continuous differentiable on each closed face, then the right-hand side of (3.3) is a $\mathcal{C}^1$-function up to the edge with continuous first derivative. Therefore, the Mellin transform with respect to the δ variable of the right-hand side exists for $Re(z) > -1$, but there is eventually a pole at $z = 0$ which corresponds to the value at $\delta = 0$. The regularity of the solution with respect to the δ-variable is determined by Lemma 3.5. We denote by

$$(4.14)\qquad \overline{\nu}_2 = \overline{\nu}(\alpha, \sqrt{\rho_1\rho_2}) \in [0, 3/2)$$

the unique solution of

$$\sqrt{\rho_1\rho_2}(1+\cos(\alpha)) B\left(\frac{3}{2} - \nu_2, \frac{3}{2} + \nu_2\right) {}_2F_1\left(\frac{1}{2} + \nu_2, \frac{1}{2} - \nu_2, 2, \frac{1+\cos(\alpha)}{2}\right) = 1,$$

where α is the angle of Figure 2, and the uniqueness follows by [17, Lemma 3.1]. The next corollary collects the results from this section and the remark after Theorem 3.6.

COROLLARY 4.2. *Let $E : S \to \mathbb{R}$ be continuous differentiable on all faces $\overline{\Delta'_j}$ (see (4.3)) of S, and denote by $\overline{u}$ the solution of (1.1).*

1. *$\overline{u}|_{\Delta'_j}$ is a $\mathcal{C}^1$-function for all j. If the boundary of Δ'_j has no points in common with an edge of S, then we have $\overline{u}|_{\overline{\Delta'_j}} \in \mathcal{C}^1(\overline{\Delta'_j})$.*

2. *If $v \in S$ is a vertex of S and Δ'_j and Δ'_k are two adjacent faces with v as a common point, we denote by $\overline{u}_v$ the localization of $\overline{u}$ to the edge $\overline{\Delta'_j} \cap \overline{\Delta'_k}$. We introduce polar coordinates and denote by $\widehat{u}_v$ the Mellin transform of $\overline{u}_v$ with respect to the r variable. Then we get*

$$\widehat{u}(z,\delta) = \widehat{u}(z,0) + O_{\delta\to 0}(\delta^{\overline{\nu}_2 - 1/2}), \quad \mathrm{Re}(z) \in (0,3), \tag{4.15}$$

where $\overline{\nu}_2$ is given by (4.14); see also Figure 3 *for the shape of the function $\overline{\nu}_2(\cdot)$.*

3. *If Δ'_j is a triangle which gets only light from faces, which can be seen from every point of Δ'_j (even if these faces are adjacent), then we have*

$$\overline{u}|_{\Delta'_j} \in \mathcal{C}^\infty(\Delta'_j) \tag{4.16}$$

if $E \in \mathcal{C}^\infty(\Delta'_j)$.

Conclusion. In section 3 we proved regularity results for the solution $\overline{u}$ of the radiosity equation (2.7) in the vicinity of an arbitrary vertex $\mathcal{V}$ of S under the assumption that the set Ω is convex. So if Ω is convex or if the contributions from distant parts of S ($\psi K(1-\phi)\overline{u}$; see (2.23)) are negligible, then Theorem 3.6, especially formula (3.34), indicates how to grade the meshes for the numerical solution in the direction toward the edges in the vicinity of a vertex. In the case of a convex set, Corollaries 3.11–3.12 further show that the solution $\overline{u}$ is a $\mathcal{C}^\infty$-function if the right-hand side of the radiosity equation is a $\mathcal{C}^\infty$-function.

In section 4, we consider the nonconvex case and get the result that we can divide the surface S in smaller triangles, such that the solution is $\mathcal{C}^1$ on these smaller triangles if the given emissivity function is a $\mathcal{C}^1$-function. This partitioning into smaller triangles is related to the discontinuity meshing; see [10, 11, 7, 9]. On these finer triangulations we can also apply the results from section 3; see Corollary 4.2(2) and 4.2(3).

Acknowledgments. The author would like to thank Professor Dr. K. Atkinson, University of Iowa, and Professor Dr. H. Multhei, Johannes Gutenberg–Universität Mainz, who made it possible for him to spend half a year at the University of Iowa, Iowa City. He would like to thank Professor Dr. K. Atkinson especially for his kind assistance during this half year and for many discussions on the radiosity equation.

REFERENCES

[1] K. Atkinson, *The planar radiosity equation and its numerical solution*, IMA J. Numer. Anal., 20 (2000), pp. 303–332.

[2] K. Atkinson and G. Chandler, *The collocation method for solving the radiosity equation for unoccluded surfaces*, J. Integral Equations Appl., 10 (1998), pp. 253–289.

[3] K. Atkinson and D. D.-K. Chien, *A fast matrix–vector multiplication method for solving the radiosity equation*, Adv. Comput. Math., 12 (2000), pp. 151–174.

[4] K. Atkinson, D. D.-K. Chien, and J. Seol, *Numerical analysis of the radiosity equation using the collocation method*, Electron. Trans. Numer. Anal., 11 (2000), pp. 94–120.

[5] F. C. Cohen and J. R. Wallace, *Radiosity and Realistic Image Synthesis*, Academic Press, Cambridge, MA, 1993.

[6] G. Drettakis, F. Durand, and C. Puech, *The visibility skeleton: A powerful and efficient multi-purpose global visibility tool*, Computer Graphics, 31 (1997), pp. 88–100.

[7] G. Drettakis, F. Durand, and C. Puech, *Fast and accurate hierarchical radiosity using global visibility*, ACM Trans. Graphics, 18 (1999), pp. 128–170.

[8] J. Elschner, *The double layer potential over polyhedral domains* I: *Solvability in weighted Sobolev spaces*, Appl. Anal., 45 (1992), pp. 117–134.

[9] S. Ghali and A. J. Stewart, *A complete treatment of D*1 *discontinuities in a discontinuity mesh*, in Proceedings of the Graphics Interface Conference, The Canadian Human Computer Communications Society, Toronto, Canada, 1996, pp. 122–131.

[10] P. Heckbert, *Radiosity in flatland*, Computer Graphics Forum, 11 (1992), pp. 181–192.

[11] P. Heckbert, *Discontinuity meshing for radiosity*, in Proceedings of the Third Eurographics Workshop on Rendering, Bristol, UK, 1992, pp. 203–226.

[12] J. R. Howell, J. Lohrengel, and R. Siegel, *Wärmeübertragung durch Strahlung, Teil* 1, Springer-Verlag, Berlin, Heidelberg, New York, 1988.

[13] P. Jeanquartier, *Transformation de Mellin et devéloppements asymptotiques*, Enseign. Math. (2), 25 (1979), pp. 285–308.

[14] W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Tables of Integral Transforms, Vol.* 1, McGraw–Hill, New York, 1954.

[15] W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions, Vol.* 1, McGraw–Hill, New York, 1955.

[16] N. Qatanani, *Lösungsverfahren und Analysis der Integralgleichung für das Hohlraum–Strahlungs–Problem*, Ph.D. thesis, Universität Stuttgart, Stuttgart, Germany, 1996.

[17] A. Rathsfeld, *Edge asymptotics for the radiosity equation over polyhedral domains*, Math. Methods Appl. Sci., 22 (1999), pp. 217–241.

[18] A. Rathsfeld, *Piecewise polynomial collocation for the double layer potential equation over polyhedral domains*, in Boundary Value Problems and Integral Equations in Nonsmooth Domains, M. Costabel, M. Dauge, and S. Nicaise, eds., Marcel Dekker, New York, 1995, pp. 219–255.

[19] H. Schon, *Effiziente Verfahren zur numerischen Lösung der Radiosity–Gleichung*, Master's thesis, Universität Karlsruhe, Karlsruhe, Germany, 1997.

[20] F. X. Sillion and C. Puech, *Radiosity and Global Illumination*, Morgan Kaufmann, San Francisco, CA, 1994.

[21] G. N. Watson and E. T. Whittaker, *A Course of Modern Analysis*, Macmillan, New York, 1948.

SIAM J. MATH. ANAL.
Vol. 33, No. 4, pp. 751–762

STABILITY OF ENTROPY SOLUTIONS TO THE CAUCHY PROBLEM FOR A CLASS OF NONLINEAR HYPERBOLIC-PARABOLIC EQUATIONS[*]

GUI-QIANG CHEN[†] AND EMMANUELE DIBENEDETTO[‡]

Abstract. Consider the Cauchy problem for the nonlinear hyperbolic-parabolic equation:

$$u_t + \frac{1}{2}\,\mathbf{a}\cdot\nabla_x u^2 = \Delta u_+ \qquad \text{for } t > 0, \tag{$*$}$$

where $\mathbf{a}$ is a constant vector and $u_+ = \max\{u, 0\}$. The equation is hyperbolic in the region $[u < 0]$ and parabolic in the region $[u > 0]$. It is shown that entropy solutions to $(*)$ that grow at most linearly as $|x| \to \infty$ are stable in a weighted $L^1(\mathbb{R}^N)$ space, which implies that the solutions are unique. The linear growth as $|x| \to \infty$ imposed on the solutions is shown to be optimal for uniqueness to hold. The same results hold if the Burgers nonlinearity $\frac{1}{2}\,\mathbf{a}u^2$ is replaced by a general flux function $\mathbf{f}(u)$, provided $\mathbf{f}'(u(x,t))$ grows in x at most linearly as $|x| \to \infty$, and/or the degenerate term u_+ is replaced by a nondecreasing, degenerate, Lipschitz continuous function $\beta(u)$ defined on $\mathbb{R}$. For more general $\beta(\cdot)$, the results continue to hold for bounded solutions.

Key words. stability, uniqueness, entropy solutions, hyperbolic-parabolic, degenerate parabolic, nonlinear equations, unbounded solutions, optimal growth, Cauchy problem

AMS subject classifications. 35K65, 35B35, 35G25, 35D99

PII. S0036141001363597

1. Introduction and results.

Consider the Cauchy problem for the nonlinear hyperbolic-parabolic equation

$$\begin{aligned} &u_t + \frac{1}{2}\,\mathbf{a}\cdot\nabla_x u^2 = \Delta u_+ \qquad \text{in } \mathcal{D}'(S_T),\\ &u(\cdot,0) = u_0 \in L^\infty_{\text{loc}}(\mathbb{R}^N), \end{aligned} \tag{1.1}$$

where $\mathbf{a}$ is a constant vector, $u_+ = \max\{u, 0\}$, the set S_T is the strip

$$S_T = \mathbb{R}^N \times (0, T] \qquad \text{for some } T > 0,$$

and the initial data are taken in the sense of $L^1_{\text{loc}}(\mathbb{R}^N)$.

The equation in (1.1) could be regarded as a prototype model for the motion of an ideal fluid filling $\mathbb{R}^N$ and exhibiting both viscous and nonviscous phases. The set $[u > 0]$ can be identified with the viscous phase, the set $[u < 0]$ is the inviscid phase, and the set $[u = 0]$ is the free boundary interface separating these phases. Accordingly, the equation in (1.1) is of mixed type; i.e., it is hyperbolic in the inviscid phase and parabolic in the viscous phase. It can also be regarded as degenerate parabolic.

The main issues relating to the Cauchy problem (1.1) are its unique solvability and the local behavior of its solutions. Both issues are relatively well understood if

[*] Received by the editors November 4, 1999; accepted for publication (in revised form) July 22, 2001; published electronically December 7, 2001.

http://www.siam.org/journals/sima/33-4/36359.html

[†] Department of Mathematics, Northwestern University, Evanston, IL 60208–2730 (gqchen@math.northwestern.edu). This author's work was partially supported by NSF grants DMS-9971793, INT-9987378, and INT-9726215.

[‡] Department of Mathematics, Vanderbilt University, Nashville, TN 60208–2730 (dibe@math.vanderbilt.edu). This author's work was partially supported by NSF grants DMS-9706388 and DMS-9708261.

one removes the hyperbolic part $\mathbf{a}\cdot\nabla u^2$, thereby obtaining a degenerate parabolic equation (see, for example, the discussion and references of [6, 7, 8]). They are equally well understood if one removes the viscosity term Δu_+. This would give the N-dimensional Burgers equation (see, for example, [3, 4, 5]).

We establish that entropy solutions of (1.1) that grow at most linearly as $|x|\to\infty$ are stable, which implies that the solutions are unique. We also show that such a growth is optimal for uniqueness to hold.

For a smooth convex function $\eta(\cdot)$, let $\mathbf{q}(\eta;u)$ denote the flux function corresponding to the entropy $\eta(u)$, i.e.,

$$q(\eta;u) = \int_0^u s\eta'(s)ds, \qquad \mathbf{q}(\eta;u)\equiv q(\eta;u)\,\mathbf{a}. \tag{1.2}$$

A function $u\in L^\infty_{\rm loc}(S_T)$ is an entropy solution of the Cauchy problem (1.1) if

$$u_+ \in L^2\big(0,T;W^{1,2}_{\rm loc}(\mathbb{R}^N)\big) \tag{1.3}$$

and if, for every convex function $\eta\in C^2(\mathbb{R})$, for any nonnegative testing function $\psi\in C^1\big([0,T);C^2_0(\mathbb{R}^N)\big)$ with $\psi|_{t\geq T}=0$,

$$\begin{aligned}&\int_{S_T}\big\{\eta(u)\,\psi_t + \mathbf{q}(\eta;u)\cdot\nabla_x\psi - \eta'(u)\nabla_x u_+\cdot\nabla_x\psi\big\}\,dxdt\\ &-\int_{S_T}\eta''(u)\,|\nabla_x u_+|^2\,\psi\,dxdt + \int_{\mathbb{R}^N}\eta(u_0)\,\psi(x,0)\,dx\geq 0.\end{aligned}\tag{1.4}$$

Entropy solutions are distributional solutions. Since $u\in L^\infty_{\rm loc}(S_T)$, an adaptation of the results of [2] implies that $u_+\in C_{\rm loc}(S_T)$.

Main Theorem. *Let u and v be two entropy solutions of* (1.1) *in the sense of* (1.2)–(1.4) *for initial data u_0 and v_0. Assume, in addition, that they satisfy the growth condition*

$$|u(x,t)|+|v(x,t)| \leq \gamma(1+|x|) \tag{1.5}$$

for almost all $(x,t)\in\mathbb{R}^N\times[0,T]$, for some positive constant γ. Then there exists a smooth, positive weight $w(x,t)$ that can be determined a priori only in terms of γ and satisfying

$$w(x,t)\,\big(1+|x|\big)\in L^1(\mathbb{R}^N), \qquad \textit{uniformly in } t\in(0,T),$$

such that

$$\int_{\mathbb{R}^N} w(x,t)\,\big|u(x,t)-v(x,t)\big|\,dx \leq \int_{\mathbb{R}^N} w(x,0)\,\big|u_0(x)-v_0(x)\big|\,dx \tag{1.6}$$

for a.e. $t\in(0,T]$.

The theorem continues to hold if the Burgers nonlinearity $\frac{1}{2}\mathbf{a}u^2$ is replaced by a more general flux function $\mathbf{f}(u)$, provided $|\mathbf{f}'(u(x,t))|$ grows in x at most linearly as $|x|\to\infty$ for any fixed $t\in[0,T)$. Also, the degenerate term u_+ can be replaced by a nondecreasing, degenerate (possibly identically zero), Lipschitz continuous function $\beta(u)$ defined in $\mathbb{R}$. These generalizations follow, mutatis mutandis, from the arguments in sections 3–7. The Lipschitz requirement on $\beta(\cdot)$ would exclude degeneracies of the type $\beta(u) = u_+^m$ for some $m > 1$. For such a $\beta(\cdot)$, the theorem continues to hold for bounded solutions.

The stability theorem and the techniques we develop here may be useful to make error estimates for various approximate solutions, especially numerical methods, and to establish similar results for more general hyperbolic-parabolic equations, which model a wide variety of phenomena, ranging from porous media flow, via flow of glaciers and sedimentation processes, to traffic flow.

In section 2, we make some remarks to show that the linear growth as $|x| \to \infty$ imposed on the solutions in our main theorem is optimal; and the solutions in general are local, and if they exist, they blow up in finite time.

In order to prove the main theorem, in section 3, we first use the definition (1.4) of entropy solutions to derive two integral inequalities for a class of C^2 entropy-entropy flux pairs, depending on a parameter $\varepsilon > 0$. These two inequalities are written for two entropy solutions u and v of (1.1) with initial data u_0 and v_0, respectively.

In section 4, we first add these two integral inequalities and choose appropriate testing functions depending upon a parameter $h > 0$. These testing functions have a mollifier-type effect and serve to handle the possible irregularity of the solutions. Then we introduce a change of variables and employ the new variables in sections 5 and 6 to transform these integral inequalities into a form suitable to study the limits as $\varepsilon \to 0$ and then $h \to 0$ in the indicated order. This will produce a further more stringent integral inequality for u and v involving testing functions still to be chosen. In section 7, such testing functions are chosen to identify the weight $w(x,t)$ and the stability result (1.6).

2. Remarks on the stability.

Distributional solutions, which are not entropy solutions, are not unique and hence are not stable. For example, nonpositive solutions of (1.1) would be solutions of the inviscid Burgers equation, for which uniqueness fails outside the class of entropy solutions [3, 4, 5].

The growth condition (1.5) is optimal for uniqueness (hence stability) to hold. For this, consider locally bounded, nonnegative solutions of (1.1) for $N = 1$, i.e.,

$$\begin{aligned} &u_t + u\,u_x = u_{xx} \quad \text{in } \mathcal{D}'(S_T), \\ &u(\cdot,0) = u_0 \geq 0 \qquad \text{and smooth in } \mathbb{R}. \end{aligned} \tag{2.1}$$

Solutions of (2.1) are smooth and positive in S_T. Set

$$v(x,t) = \int_\alpha^x u(y,t)\,dy \qquad \text{for some } \alpha \in \mathbb{R}.$$

Then $u = v_x$, and u is a solution of (2.1) if and only if v is a classical solution of

$$\begin{aligned} &v_t + \frac{1}{2}(v_x)^2 - v_{xx} = c_\alpha(t), \\ &v(x,0) = \int_\alpha^x u_0(y)\,dy, \end{aligned} \tag{2.1$_v$}$$

where

$$c_\alpha(t) = \frac{1}{2}u^2(\alpha,t) - u_x(\alpha,t).$$

The Hopf-like transformation

$$w(x,t) \;=\; \exp\left\{-\frac{1}{2}\,v \,+\, \frac{1}{2}\int_0^t c_\alpha(\tau)\,d\tau\right\}$$

transforms $(2.1)_v$ into the equivalent formulation

$$(2.1)_w \qquad \begin{aligned} w_t \,-\, w_{xx} \;&=\; 0 \qquad \text{in } \; S_T, \\ w(x,0) \;&=\; \exp\left\{-\frac{1}{2}\,v(x,0)\right\}. \end{aligned}$$

Therefore, well-posedness for (2.1) is equivalent to well-posedness for $(2.1)_w$. The latter is well-posed if and only if

$$w(x,t) \;\leq\; \exp\left\{C\,\left(1 \,+\, x^2\right)\right\}$$

for some positive constant C and for all $t \in [0,T]$ (see [1, Chap. 5]). The constant C and the time T are linked by $4CT < 1$. By the Tychonov counterexample [1, p. 237], a faster growth would not guarantee uniqueness.

In terms of $u = -2(e^w)_x$, this implies that the stability theorem would be false for solutions growing faster than linearly as $|x| \to \infty$. This also implies that solutions of (1.1) are, in general, not global in time.

If the initial datum u_0 is bounded in $\mathbb{R}^N$ and in $BV(\mathbb{R}^N)$, then solutions of (1.1) can be constructed as in Volpert–Hudjaev [7, 8]. For such initial data, the authors also proved uniqueness. If the initial datum u_0 is bounded in $\mathbb{R}^N$ but not necessarily regular (i.e., not in $BV(\mathbb{R}^N)$), the existence of entropy solutions can be established as in Kruzhkov [3, 4], as indicated also in [6]. The same construction also implies $\nabla u_+ \in L^2_{\mathrm{loc}}(S_T)$. Attention to the uniqueness problem for parabolic-hyperbolic equations such as (1.1) has been drawn in [6].

If the initial datum is unbounded, then, by the previous remarks, solutions in general are local, and if they exist, they blow up in finite time. An existence theorem would have to involve an estimation of the blow-up time.

3. Proof of the stability theorem (i). In this section, we first use the definition (1.4) of entropy solutions to derive two integral inequalities for a class of C^2 entropy-entropy flux pairs, depending on a parameter $\epsilon > 0$.

To choose suitable entropy functions $\eta(u)$ in (1.4), introduce the regularizations of the Heaviside function:

$$H_\varepsilon(s) \equiv \begin{cases} 1 & \text{if } \; s > \varepsilon, \\ \sin\left(\dfrac{\pi}{2\varepsilon}\,s\right) & \text{if } \; |s| \leq \varepsilon, \\ -1 & \text{if } \; s < -\varepsilon. \end{cases}$$

Then, for each $k \in \mathbb{R}$, the functions

$$(3.1) \qquad u \longrightarrow \eta_\varepsilon(u-k) \equiv \int_0^{u-k} H_\varepsilon(s)\,ds, \qquad 0 < \varepsilon \ll 1,$$

are convex and of class C^2 in $\mathbb{R}$. Moreover,

$$(3.1)' \qquad \eta_\varepsilon(u-k) \longrightarrow |u-k| \qquad \text{as } \; \varepsilon \,\to\, 0.$$

For the corresponding flux functions defined in (1.2), we set

$$\mathbf{Q}_\varepsilon(u;k) \,=\, \mathbf{q}(\eta_\varepsilon;u) \,\equiv\, \left(\int_k^u s\,H_\varepsilon(s-k)\,ds\right)\mathbf{a}. \tag{3.2}$$

One verifies that

$$\mathbf{Q}_\varepsilon(u;k) \longrightarrow \frac{1}{2}\left|u^2-k^2\right|\mathbf{a} \qquad \text{as} \quad \varepsilon \to 0. \tag{3.2$'$}$$

In (1.4) we choose nonnegative testing functions $(x,t;y,\tau) \to \psi(x,t;y,\tau)$ depending upon two pairs of variables (x,t) and (y,τ) such that

$$\begin{aligned} &(x,t) \,\to\, \psi(x,t;y,\tau) \,\in\, C^1\big([0,T);C_0^2(\mathbb{R}^N)\big) \quad \text{ uniformly in } (y,\tau),\\ &(y,\tau) \,\to\, \psi(x,t;y,\tau) \,\in\, C^1\big([0,T);C_0^2(\mathbb{R}^N)\big) \quad \text{ uniformly in } (x,t). \end{aligned}$$

We also require that $\psi(x,t;y,\tau)=0$ for $t \ge T$ or $\tau \ge T$. With such a choice, (1.4) can be written interchangeably in terms of either pair of variables.

In (1.4), written in terms of (x,t), choose the entropy function

$$(x,t) \longrightarrow \eta_\varepsilon\big(u(x,t)-v(y,\tau)\big), \qquad (y,\tau)\in S_T \text{ fixed.} \tag{3.1$_{(x,t)}$}$$

For the choice of $k=v(y,\tau)$, the flux function introduced in (3.2) takes the form

$$\mathbf{Q}_\varepsilon\big(u(x,t);v(y,\tau)\big) \,=\, \left(\int_{v(y,\tau)}^{u(x,t)} s\,H_\varepsilon\big(s-v(y,\tau)\big)\,ds\right)\mathbf{a}. \tag{3.2$_{(x,t)}$}$$

We put these choices in (1.4) and then integrate over S_T in $dyd\tau$ to obtain

(3.3)$_{(x,t)}$

$$\begin{aligned} \int_{S_T}\int_{S_T}\Big\{\, &\eta_\varepsilon\big(u(x,t)-v(y,\tau)\big)\psi_t + \mathbf{Q}_\varepsilon\big(u(x,t);v(y,\tau)\big)\cdot\nabla_x\psi\\ &-H_\varepsilon\big(u(x,t)-v(y,\tau)\big)\nabla_x u_+(x,t)\cdot\nabla_x\psi\\ &-H'_\varepsilon\big(u(x,t)-v(y,\tau)\big)|\nabla_x u_+(x,t)|^2\,\psi\Big\}dxdt\,dyd\tau\\ &+\int_{S_T}\int_{\mathbb{R}^N}\eta_\varepsilon\big(u_0(x)-v(y,\tau)\big)\,\psi(x,0;y,\tau)\,dx\,dyd\tau \,\ge\, 0. \end{aligned}$$

Next we write the weak formulation (1.1) for the solution $(y,\tau)\to v(y,\tau)$, and in the resulting expression we take the entropy function

$$(y,\tau) \longrightarrow \eta_\varepsilon\big(v(y,\tau)-u(x,t)\big), \qquad (x,t)\in S_T \text{ fixed,} \tag{3.1$_{(y,\tau)}$}$$

and the corresponding flux function

$$\mathbf{Q}_\varepsilon\big(v(y,\tau);u(x,t)\big) = \left(\int_{u(x,t)}^{v(y,\tau)} s\,H_\varepsilon\big(s-u(x,t)\big)\,ds\right)\mathbf{a}. \tag{3.2$_{(y,\tau)}$}$$

Integrating over S_T in $dxdt$ yields

$(3.3)_{(y,\tau)}$
$$\int_{S_T}\int_{S_T}\Big\{\eta_\varepsilon\big(v(y,\tau)-u(x,t)\big)\psi_\tau + \mathbf{Q}_\varepsilon\big(v(y,\tau);u(x,t)\big)\cdot\nabla_y\psi$$
$$-H_\varepsilon\big(v(y,\tau)-u(x,t)\big)\nabla_y v_+(y,\tau)\cdot\nabla_y\psi$$
$$-H'_\varepsilon\big(v(y,\tau)-u(x,t)\big)|\nabla_y v_+(y,\tau)|^2\,\psi\Big\}dxdt\,dyd\tau$$
$$+\int_{S_T}\int_{\mathbb{R}^N}\eta_\varepsilon\big(v_0(y)-u(x,t)\big)\,\psi(x,t;y,0)dy\,dxdt \ \ge\ 0.$$

4. Proof of the stability theorem (ii). In this section, we first add the inequalities $(3.3)_{(x,t)}$–$(3.3)_{(y,\tau)}$ and choose appropriate testing functions depending upon a parameter $h>0$, and then we introduce a change of variables to transform these integral inequalities into a form suitable to study the limits as ϵ and h tend to zero.

We now add the inequalities $(3.3)_{(x,t)}$–$(3.3)_{(y,\tau)}$ and use the facts that η_ε and η''_ε are even functions and η'_ε is odd to obtain

$$I_{1,\varepsilon} + I_{2,\varepsilon} + I_{3,\varepsilon} \ge 0, \tag{4.1}$$

where we have set

$$I_{1,\varepsilon} = \int_{S_T}\int_{S_T}\Big\{\ \eta_\varepsilon\big(u(x,t)-v(y,\tau)\big)\,(\psi_t + \psi_\tau)$$
$$+\mathbf{Q}_\varepsilon\big(u(x,t);v(y,\tau)\big)\cdot\nabla_x\psi \tag{4.2}$$
$$+\mathbf{Q}_\varepsilon\big(v(y,\tau);u(x,t)\big)\cdot\nabla_y\psi\Big\}dxdt\,dyd\tau,$$

$$I_{2,\varepsilon} = -\int_{S_T}\int_{S_T}\Big\{H_\varepsilon\ \big(u(x,t)-v(y,\tau)\big)$$
$$\times\big[\nabla_x u_+(x,t)\cdot\nabla_x\psi\ -\ \nabla_y v_+(y,\tau)\cdot\nabla_y\psi\big]$$
$$+\ H'_\varepsilon\big(u(x,t)-v(y,\tau)\big) \tag{4.3}$$
$$\times\big[|\nabla_x u_+(x,t)|^2 + |\nabla_y v_+(y,\tau)|^2\big]\psi\Big\}dxdt\,dyd\tau,$$

and

$$I_{3,\varepsilon} = \int_{S_T}\int_{\mathbb{R}^N}\eta_\varepsilon\big(u_0(x)-v(y,\tau)\big)\psi(x,0;y,\tau)dx\,dyd\tau$$
$$+\int_{S_T}\int_{\mathbb{R}^N}\eta_\varepsilon\big(v_0(y)-u(x,t)\big)\psi(x,t;y,0)dy\,dxdt. \tag{4.4}$$

Next we choose the function $(x,t;y,\tau)\to\psi(x,t;y,\tau)$ of the form

$$\psi(x,t;y,\tau) = \varphi\left(\frac{1}{2}(x+y);\frac{1}{2}(t+\tau)\right) j_h\left(\frac{1}{2}(x-y);\frac{1}{2}(t-\tau)\right),$$

where $\varphi(\cdot;\cdot) \in C_0^\infty(S_T)$ is nonnegative and

$$j_h\left(\frac{1}{2}(x-y);\frac{1}{2}(t-\tau)\right) = \omega_h\left(\frac{1}{2}|x-y|\right)\omega_h\left(\frac{1}{2}(t-\tau)\right).$$

Here ω denotes the standard, symmetric mollifying kernel in $\mathbb{R}$, and

$$\omega_h(s) = \frac{1}{h}\,\omega\left(\frac{s}{h}\right), \quad \text{and} \quad \omega_h(s) = 0 \quad \text{for } |s| \geq h.$$

Consider the change of variables:

$$\begin{array}{llll} \xi = \frac{1}{2}(x+y), & \zeta = \frac{1}{2}(x-y), & s = \frac{1}{2}(t+\tau), & \sigma = \frac{1}{2}(t-\tau), \\ x = \xi+\zeta, & y = \xi-\zeta, & t = s+\sigma, & \tau = s-\sigma, \end{array} \tag{4.5}$$

whose Jacobian is 4. As $(x,t;y,\tau)$ range over $S_T \times S_T$, the new variables (ξ, s) range over S_T and (ζ,σ) range over S_T', where

$$S_T' = \mathbb{R}^N \times \left(-\frac{1}{2}T\,,\,\frac{1}{2}T\right).$$

Therefore, the change of variables in (4.5) maps $S_T \times S_T$ into $S_T \times S_T'$. In terms of the new variables, we compute

$$\begin{aligned} \psi(x,t;y,\tau) &= \varphi(\xi;s)\,j_h(\zeta;\sigma), \\ \nabla_x\psi &= \frac{1}{2}\left\{\nabla_\xi\varphi\,j_h + \varphi\,\nabla_\zeta j_h\right\}, \\ \nabla_y\psi &= \frac{1}{2}\left\{\nabla_\xi\varphi\,j_h - \varphi\,\nabla_\zeta j_h\right\}, \\ \nabla_x u_+(x,t) &= \frac{1}{2}\left\{\nabla_\xi u_+(\xi+\zeta\,,\,s+\sigma) + \nabla_\zeta u_+(\xi+\zeta\,,\,s+\sigma)\right\} \\ &= \nabla_\xi u_+(\xi+\zeta\,,\,s+\sigma), \\ \nabla_y v_+(y,\tau) &= \frac{1}{2}\left\{\nabla_\xi v_+(\xi-\zeta\,,\,s-\sigma) + \nabla_\xi v_+(\xi-\zeta\,,\,s-\sigma)\right\} \\ &= \nabla_\xi v_+(\xi-\zeta\,,\,s-\sigma), \\ \nabla_x j_h + \nabla_y j_h &= 0, \qquad j_{h,t} + j_{h,\tau} = 0, \\ \nabla_x\psi + \nabla_y\psi &= \nabla_\xi\varphi\,j_h, \qquad \psi_t + \psi_\tau = \varphi_s\,j_h, \\ \varphi_{\zeta_i} &= 0, \qquad j_{h,\xi_i} = 0, \quad i = 1,2,\ldots,N. \end{aligned}$$

In view of the dependence of u upon $(\xi+\zeta)$ and of v upon $(\xi-\zeta)$,

$$\begin{aligned} \nabla_\xi u_+(\xi+\zeta\,,\,s-\sigma) &= \nabla_\zeta u_+(\xi+\zeta\,,\,s-\sigma), \\ \nabla_\xi v_+(\xi-\zeta\,,\,s-\sigma) &= -\nabla_\zeta v_+(\xi-\zeta\,,\,s-\sigma). \end{aligned} \tag{4.6}$$

Next, using these new variables, we transform the various integrals in (4.1).

5. Transformation and limits of $I_{2,\varepsilon}$. We first use the new variables introduced in section 4 to transform $I_{2,\varepsilon}$ into a form suitable to study the limits as $\varepsilon \to 0$ and then $h \to 0$ in the indicated order, and we then estimate the limits.

With the indicated choices and change of variables, we have

$$\begin{aligned}
-\frac{1}{2}I_{2,\varepsilon} &= \int_{S'_T}\int_{S_T} H_\varepsilon\big(u(\xi+\zeta,s+\sigma)-v(\xi-\zeta,s-\sigma)\big)\\
&\quad\times\Big\{\,\nabla_\xi u_+(\xi+\zeta,s+\sigma)\cdot(\nabla_\xi\varphi\, j_h+\varphi\nabla_\zeta j_h)\\
&\qquad -\nabla_\xi v_+(\xi-\zeta,s-\sigma)\,(\nabla_\xi\varphi\, j_h-\varphi\nabla_\zeta j_h)\,\Big\}d\xi ds\,d\zeta d\sigma\\
&\qquad +2\int_{S'_T}\int_{S_T} H'_\varepsilon\big(u(\xi+\zeta,s+\sigma)-v(\xi-\zeta,s-\sigma)\big)\\
&\qquad \times\left\{|\nabla_\xi u_+|^2\,+\,|\nabla_\xi v_+|^2\right\}\,\varphi\, j_h\, d\xi ds\,d\zeta d\sigma\\
&= \int_{S'_T}\int_{S_T} H_\varepsilon(u-v)\nabla_\xi(u_+-v_+)\cdot\nabla_\xi\varphi\, j_h\, d\xi ds\,d\zeta d\sigma\\
&\quad +\int_{S'_T}\int_{S_T} H_\varepsilon(u-v)\nabla_\xi(u_++v_+)\cdot\varphi\,\nabla_\zeta j_h\, d\xi ds\,d\zeta d\sigma\\
&\quad +2\int_{S'_T}\int_{S_T} H'_\varepsilon\big(u(\xi+\zeta,s+\sigma)-v(\xi-\zeta,s-\sigma)\big)\\
&\quad \times\left\{|\nabla_\xi u_+|^2\,+\,|\nabla_\xi v_+|^2\right\}\,\varphi\, j_h\, d\xi ds\,d\zeta d\sigma\\
&= I^{(1)}_{2,\varepsilon}\,+\,I^{(2)}_{2,\varepsilon}\,+\,I^{(3)}_{2,\varepsilon}.
\end{aligned}$$

In transforming these integrals, we make use of the integrability of ∇u_+. In particular, $|\nabla u_+|$ vanishes a.e. on the set $[u\le 0]$, and a similar fact holds for v_+. Then, for a.e. $(\xi,s;\zeta,\sigma)\in S_T\times S'_T$, we write

$$\begin{aligned}
H_\varepsilon(u-v)\nabla_\xi(u_+-v_+) &= H_\varepsilon(u_+-v_+)\nabla_\xi(u_+-v_+)\\
&\quad +\nabla_\xi u_+\left\{H_\varepsilon(u_+-v)-H_\varepsilon(u_+-v_+)\right\}\\
&\quad +\nabla_\xi v_+\left\{H_\varepsilon(u_+-v_+)-H_\varepsilon(u-v_+)\right\}.
\end{aligned}$$

As $\varepsilon\to 0$, the last two terms tend to zero a.e. on every compact subset of $S_T\times S'_T$, and their modulus is dominated, uniformly in ε, by a locally integrable function. Thus, when they are put in the expression of $I^{(1)}_{2,\varepsilon}$ and after we take the limit as $\varepsilon\to 0$, they give no contribution. This process in $I^{(1)}_{2,\varepsilon}$ gives

$$\begin{aligned}
\lim_{\varepsilon\to 0} I^{(1)}_{2,\varepsilon} &= \lim_{\varepsilon\to 0}\int_{S'_T}\int_{S_T}\nabla_\xi\left(\int_0^{u_+-v_+} H_\varepsilon(\theta)d\theta\right)\cdot\nabla_\xi\varphi\, j_h\, d\xi ds\,d\zeta d\sigma\\
&= -\lim_{\varepsilon\to 0}\int_{S'_T}\int_{S_T}\eta_\varepsilon(u_+-v_+)\,\Delta_\xi\varphi\, j_h\, d\xi ds\,d\zeta d\sigma\\
&= -\int_{S'_T}\int_{S_T}|u_+-v_+|\,\Delta_\xi\varphi\, j_h\, d\xi ds\,d\zeta d\sigma.
\end{aligned}$$

In transforming $I^{(2)}_{2,\varepsilon}$, we first assume that u_+ and v_+ are regular, and we proceed formally. By a repeated formal integration by parts,

$$\begin{aligned}
I^{(2)}_{2,\varepsilon} &= -\int_{S'_T}\int_{S_T} H'_\varepsilon(u_+-v_+)\,\nabla_\xi(u_++v_+)\cdot\nabla_\zeta(u_+-v_+)\,\varphi\, j_h\, d\xi ds\, d\zeta d\sigma \\
&\quad -\int_{S'_T}\int_{S_T} H_\varepsilon(u_+-v_+)\,\mathrm{div}_\zeta\nabla_\xi(u_++v_+)\,\varphi\, j_h\, d\xi ds\, d\zeta d\sigma \\
&= -\int_{S'_T}\int_{S_T} H'_\varepsilon(u_+-v_+)\Big\{\nabla_\xi(u_++v_+)\cdot\nabla_\zeta(u_+-v_+) \\
&\qquad\qquad -\nabla_\xi(u_+-v_+)\cdot\nabla_\zeta(u_++v_+)\Big\}\,\varphi\, j_h\, d\xi ds\, d\zeta d\sigma \\
&\quad +\int_{S'_T}\int_{S_T} H_\varepsilon(u_+-v_+)\,\nabla_\zeta(u_++v_+)\cdot\nabla_\xi\varphi\, j_h\, d\xi ds\, d\zeta d\sigma.
\end{aligned}\tag{5.1}$$

From this, taking into account the differentiation formulae (4.6) and performing a further integration by parts, we have

$$\begin{aligned}
I^{(2)}_{2,\varepsilon} &= \int_{S'_T}\int_{S_T} H'_\varepsilon(u_+-v_+)\big\{|\nabla_\xi(u_+-v_+)|^2 \\
&\qquad\qquad -|\nabla_\xi(u_++v_+)|^2\big\}\,\varphi\, j_h\, d\xi ds\, d\zeta d\sigma \\
&\quad -\int_{S'_T}\int_{S_T}\left(\int_0^{(u_+-v_+)} H_\varepsilon(\theta)d\theta\right)\Delta_\xi\varphi\, j_h\, d\xi ds\, d\zeta d\sigma.
\end{aligned}\tag{5.2}$$

These calculations can be made rigorous by the following procedure. Denote by $u_{+,\nu}$ and $v_{+,\nu}$ the mollifications of u_+ and v_+ with respect to the variables ξ and ζ. Then

$$I^{(2)}_{2,\varepsilon} = o_\nu(1) + \int_{S'_T}\int_{S_T} H_\varepsilon(u_{+,\nu}-v_{+,\nu})\nabla_\xi(u_{+,\nu}+v_{+,\nu})\cdot\varphi\,\nabla_\zeta j_h\, d\xi ds\, d\zeta d\sigma,$$

where, for $\varepsilon > 0$ fixed, $o_\nu(1) \to 0$ as $\nu \to 0$. We perform integrations by parts in the integrals involving $u_{+,\nu}$ and $v_{+,\nu}$ to arrive at a formula analogous to (5.2). We then let $\nu \to 0$ to obtain (5.2), the various limits being justified, since ∇u_+ and ∇v_+ are in $L^2_{\mathrm{loc}}(S_T)$. Finally, letting $\varepsilon \to 0$ gives

$$\begin{aligned}
\liminf_{\varepsilon\to 0} I^{(2)}_{2,\varepsilon} &= -\int_{S'_T}\int_{S_T} |u_+-v_+|\,\Delta_\xi\varphi\, j_h\, d\xi ds\, d\zeta d\sigma \\
&\quad -4\liminf_{\varepsilon\to 0}\int_{S'_T}\int_{S_T} H'_\varepsilon(u_+-v_+)\,\nabla_\xi u_+\cdot\nabla_\xi v_+\,\varphi\, j_h\, d\xi ds\, d\zeta d\sigma.
\end{aligned}$$

We now combine these calculations in the expression of $I_{2,\varepsilon}$ and let $\varepsilon \to 0$ to obtain

$$\begin{aligned}
\limsup_{\varepsilon\to 0} I_{2,\varepsilon} &= 4\int_{S'_T}\int_{S_T} |u_+-v_+|\,\Delta_\xi\varphi\, j_h\, d\xi ds\, d\zeta d\sigma \\
&\quad -4\limsup_{\varepsilon\to 0}\int_{S'_T}\int_{S_T} H'_\varepsilon(u_+-v_+)|\nabla_\xi(u_+-v_+)|^2\varphi\, j_h\, d\xi ds\, d\zeta d\sigma \\
&\le 4\int_{S'_T}\int_{S_T} |u_+-v_+|\,\Delta_\xi\varphi\, j_h\, d\xi ds\, d\zeta d\sigma
\end{aligned}$$

since $\varphi \geq 0$ and $j_h \geq 0$. Finally, we let $h \searrow 0$ by following the same arguments as in Kruzhkov [3, 4] to obtain

$$\lim_{h \to 0} \limsup_{\varepsilon \to 0} I_{2,\varepsilon} \leq 4 \int_{S_T} \left|u_+(x,t) - v_+(x,t)\right| \Delta_x \varphi(x,t)\, dx\, dt.$$

6. Transformation and limits of $I_{1,\varepsilon}$ and $I_{3,\varepsilon}$. Now we continue to perform the change of variables to transform $I_{1,\varepsilon}$ and $I_{3,\varepsilon}$ into a form suitable to study the limits as $\varepsilon \to 0$ and then $h \to \varepsilon$.

In transforming $I_{1,\varepsilon}$, we use the definitions (3.1)–(3.1)$'$ of the entropy and the definitions (3.2)–(3.2)$'$ of the corresponding flux functions. Taking into account (4.5), we compute

$$\begin{aligned}\lim_{\varepsilon \to 0} I_{1,\varepsilon} = 4 \int_{S'_T} \int_{S_T} \Big\{ &|u(\xi+\zeta, s+\sigma) - v(\xi-\zeta, s-\sigma)|\, \varphi_s(\xi,s)\, j_h(\zeta,\sigma) \\ &+ \frac{1}{2}\left|u^2(\xi+\zeta, s+\sigma) - v^2(\xi-\zeta, s-\sigma)\right| \mathbf{a} \cdot \nabla_\xi \varphi(\xi,s)\, j_h(\zeta,\sigma) \Big\} d\xi ds\, d\zeta d\sigma.\end{aligned}$$

Letting now $h \to 0$, we find

$$\begin{aligned}\lim_{\varepsilon \to 0} I_{1,\varepsilon} = 4 \int_{S_T} \Big\{ &|u(x,t) - v(x,t)| \varphi_t(x,t) \\ &+ \frac{1}{2}\left|u^2(x,t) - v^2(x,t)\right| \mathbf{a} \cdot \nabla_x \varphi(x,t) \Big\}\, dx\, dt.\end{aligned}$$

In transforming $I_{3,\varepsilon}$, we perform the change of variables (4.5) involving only the space variables, whereas the time variables are left unchanged. The Jacobian of the transformation is 2. Analogous arguments and limiting processes yield

$$\begin{aligned}\lim\nolimits_{\varepsilon \to 0} I_{3,\varepsilon} = 2 &\int_{S_T} \int_{\mathbb{R}^N} |u_0(\xi+\zeta) - v(\xi-\zeta,\tau)| \varphi\left(\xi, \frac{1}{2}\tau\right) j_h\left(\zeta, -\frac{1}{2}\tau\right) d\xi\, d\zeta d\tau \\ &+2 \int_{S_T} \int_{\mathbb{R}^N} \left|v_0(\xi-\zeta) - u(\xi+\zeta,t)\right| \varphi\left(\xi, \frac{1}{2}t\right) j_h\left(\zeta, \frac{1}{2}t\right) d\xi\, d\zeta dt.\end{aligned}$$

Now letting $h \to 0$ gives

$$\lim_{h \to 0} \lim_{\varepsilon \to 0} I_{3,\varepsilon} = 4 \int_{\mathbb{R}^N} \left|u_0(x) - v_0(x)\right| \varphi(x,0)\, dx.$$

7. Proof of the stability theorem (iii). By combining these calculations in (4.1) and after taking the limit first for $\varepsilon \searrow 0$ and then for $h \searrow 0$, we arrive at

$$\begin{aligned}\int_{S_T} \Big\{ &\left|u(x,t) - v(x,t)\right| \varphi_t(x,t) + \frac{1}{2}\left|u^2(x,t) - v^2(x,t)\right| \mathbf{a} \cdot \nabla_x \varphi(x,t) \\ &+ \left|u_+(x,t) - v_+(x,t)\right| \Delta_x \varphi(x,t) \Big\} dxdt \\ &+ \int_{\mathbb{R}^N} \left|u_0(x) - v_0(x)\right| \varphi(x,0) dx \geq 0.\end{aligned} \tag{7.1}$$

In this more stringent integral inequality, the nonnegative testing function φ is still to be chosen.

In this section, we choose the testing function to identify the weight $w(x,t)$ and the stability result (1.6).

First we rewrite (7.1) in the form

$$\begin{aligned} &\int_{S_T} \big|u(x,\tau) - v(x,\tau)\big| \big\{\varphi_\tau + \mathbf{A}\cdot\nabla_x\varphi + b\,\Delta_x\varphi\big\}\,dx\,d\tau \\ &\qquad\qquad + \int_{\mathbb{R}^N} \big|u_0(x) - v_0(x)\big|\varphi(x,0)\,dx \,\geq\, 0, \end{aligned} \tag{7.2}$$

where we have set

$$\begin{aligned} \mathbf{A}(x,\tau) &\equiv \frac{1}{2}\frac{\big|u^2(x,\tau) - v^2(x,\tau)\big|}{|u(x,\tau) - v(x,\tau)|}\,\mathbf{a} \,=\, \frac{1}{2}\,|u(x,\tau) + v(x,\tau)|\,\mathbf{a}, \\ b(x,\tau) &\equiv \frac{\big|u_+(x,\tau) - v_+(x,\tau)\big|}{\big|u(x,\tau) - v(x,\tau)\big|} \end{aligned}$$

if $u \neq v$, and $A = b = 0$ if $u = v$. In (7.2) we choose the testing functions

$$\varphi_{\varepsilon,\delta}(x,\tau) \,=\, h_\delta(\tau)\,w(x,\tau)\,\zeta(x).$$

Here, for $0 < \delta \ll 1$, we have set

$$h_\delta(\tau) \,=\, \frac{1}{\delta}\int_{\tau - t}^{\infty} \omega\left(\frac{s}{\delta}\right) ds, \qquad 0 < t \leq T - \delta,$$

where $\omega(\cdot)$ is the standard, symmetric mollifier in $\mathbb{R}$. Moreover,

$$w(x,\tau) \,=\, e^{-\,(1 + \lambda_1\,\tau)\,|x|^2 \,-\, \lambda_2\,\tau}$$

for positive constants λ_1 and λ_2 to be chosen. Finally, $x \to \zeta(x)$ is a standard, nonnegative cutoff function in the ball $\{|x| < 2R\}$, satisfying

$$\begin{cases} \zeta(x) = 1 & \text{for } \ |x| < R, \\ |\nabla\zeta| \,\leq\, \dfrac{1}{R} & \text{for all } \ |x| < 2R, \\ |\Delta\zeta| \,\leq\, \dfrac{\text{const}}{R^2} & \text{for all } \ |x| < 2R. \end{cases}$$

These testing functions are admissible since both h_δ and $w\zeta$ are nonnegative and regular, and, by the properties of the mollifiers, $h_\delta(\tau) = 0$ for $\tau \geq t + \delta$. We compute

$$\begin{aligned} w_\tau(x,\tau) &\,=\, -\big(\lambda_2 \,+\, \lambda_1\,|x|^2\big)\; w(x,\tau), \\ \nabla_x w(x,\tau) &\,=\, -2\;(1 \,+ \lambda_1\,\tau)\;x\,w(x,\tau), \\ \Delta_x w(x,\tau) &\,=\; 2\,(1 \,+\, \lambda_1\,\tau)\,\big\{2\,(1 \,+\, \lambda_1\,\tau)\,|x|^2 \,-\, N\big\}\,w(x,\tau). \end{aligned}$$

By the structure of $x \to w(x,\tau)$ and $\zeta(x)$,

$$\lim_{R\to\infty}\int_{\{R<|x|<2R\}} \{w\,(\zeta + |\nabla\zeta| + |\Delta\zeta|) + |\nabla_x w\cdot\nabla\zeta|\}\,dx \,=\, 0,$$

uniformly in $\tau \in (0,T)$. Putting this testing function in (7.2) and letting $R \to \infty$ give

$$\begin{aligned} &- \int_{S_T} |u(x,\tau) - v(x,\tau)| \, \frac{1}{\delta} \, \omega\left(\frac{\tau - t}{\delta}\right) w(x,\tau)\, dx\, d\tau \\ (7.3) \qquad &+ \int_{S_T} |u(x,\tau) - v(x,\tau)| \, h_\delta(\tau) \left\{ w_\tau + \mathbf{A} \cdot \nabla_x w + b\Delta_x w \right\} dx\, d\tau \\ &+ \int_{\mathbb{R}^N} |u_0(x) - v_0(x)| \, h_\delta(0)\, w(x,0)\, dx \;\geq\; 0. \end{aligned}$$

The growth condition (1.5) implies that there exists a positive constant C, depending only upon γ, such that

$$\mathbf{A} \cdot x \;\leq\; C\left(1 + |x|^2\right) \qquad \text{for all } \; x \in \mathbb{R}^N.$$

Then

$$w_\tau + \mathbf{A} \cdot \nabla_x w + b\Delta_x w \leq w\left\{ C - \lambda_2 + \left[C + 4(1 + \lambda_1 \tau)^2 - \lambda_1 \right] |x|^2 \right\}.$$

Choose $\lambda_2 = C$ and $T_0 = \lambda_1^{-1}$. Then select λ_1 from

$$C \,+\, 4\,(1 \,+\, \lambda_1\,\tau)^2 \,-\, \lambda_1 \;\leq\; C \,+\, 16 \,-\, \lambda_1 \;\leq\; 0.$$

For such choices, we discard the second integral in (7.3) and let $\delta \to 0$ to obtain

$$\int_{\mathbb{R}^N} w(x,t) |u(x,t) - v(x,t)|\, dx \;\leq\; \int_{\mathbb{R}^N} w(x,0) |u_0 - v_0|\, dx$$

for a.e. $t \in (0, T_0)$. Since the weight w depends only upon γ, the process can be repeated to exhaust, in a finite number of steps, the time interval of existence.

In the case of general nonlinearities $\mathbf{f}(u)$ and $\beta(u)$, the weight w can be constructed depending only on the Lipschitz constant of $\beta(u)$ and $\sup_{(x,t)\in S_T} |\mathbf{f}'(u(x,t))|/(1 + |x|)$. □

Note added in proof. After we submitted this paper, we were aware of Carrillo's paper [9] dealing with bounded solutions for a similar problem in bounded domains.

REFERENCES

[1] E. DiBenedetto, *Partial Differential Equations*, Birkhäuser Boston, Boston, 1995.

[2] E. DiBenedetto, *Continuity of weak solutions to certain singular parabolic equations*, Ann. Mat. Pura Appl. (4), 130 (1982), pp. 131–176.

[3] S. Kruzhkov, *First order quasilinear equations with several space variables*, Mat. Sb., 123 (1970), pp. 228–255 (in Russian).

[4] S. Kruzhkov, *First order quasilinear equations in several independent variables*, Math. USSR Sb., 10 (1970), pp. 217–243.

[5] P. D. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, CBMS Reg. Conf. Ser. Appl. Math. 11, SIAM, Philadelphia, 1973.

[6] P.-L. Lions, B. Perthame, and E. Tadmor, *A kinetic formulation of multidimensional scalar conservation laws and related equations*, J. Amer. Math. Soc., 7 (1994), pp. 169–191.

[7] A. I. Volpert and S. I. Hudjaev, *Cauchy problem for degenerate, second order quasilinear parabolic equations*, Mat. Sb., 78 (1969), pp. 374–396 (in Russian).

[8] A. I. Volpert and S. I. Hudjaev, *Cauchy's problem for degenerate second order quasilinear parabolic equations*, Math. USSR Sb., 7 (1970), pp. 365–387.

[9] J. Carrillo, *Entropy Solutions for Nonlinear Degenerate Problems*, Preprint, Universidad Complutense de Madrid, Madrid, Spain, 1999.

SIAM J. Math. Anal.
Vol. 33, No. 4, pp. 763–785

ON EXISTENCE OF GLOBAL SOLUTIONS AND BLOW-UP TO A SYSTEM OF REACTION-DIFFUSION EQUATIONS MODELLING CHEMOTAXIS[*]

YIN YANG[†], HUA CHEN[†], AND WEIAN LIU[†]

Abstract. In this paper we investigate the properties of the solutions for some general reaction-diffusion systems due to Othmer–Stevens which arise in modelling chemotaxis, and we prove some results about collapse and finite-time action of certain local modifications of the environment.

Key words. chemotaxis, blow-up, collapse, subsuper solution

AMS subject classifications. 35K50, 35M10, 35R25, 92C45

PII. S0036141000337796

1. Introduction.

In biology, it is very important to investigate the movement of some cells or organisms in some given biological system (cf. [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]). The mechanism of communication between cells or organisms depends on the different ways they interact. In many biological systems, the movement occurs in response to a diffusible substance or otherwise transported signal. Other systems are modelled by so-called short-range interactions due to local modifications of the environment such as the production and release of nutrients. In this case, dispersal is not simply one of simple diffusion but rather one of correlated or reinforced random walks. In order to understand how the movement rules are affected by the effect of the chemo-attractant, Othmer and Stevens introduced in [4] several general classes of partial differential equations. In one of their models, they considered a master equation, i.e., barrier, and nearest neighbor lattice model. Following a limiting process, the model is described by the following system of partial differential equations:

$$\begin{cases} \dfrac{\partial p}{\partial t} = D\nabla\cdot\left(p\nabla\left(\ln\left(\dfrac{p}{w}\right)\right)\right) & \\ \dfrac{\partial w}{\partial t} = F(p,w) & \text{for } x\in\Omega,\quad t>0, \\ p\nabla\left(\ln\left(\dfrac{p}{w}\right)\right)\cdot\vec{n} = 0 & \text{for } x\in\partial\Omega,\quad t>0, \\ p(x,0) = p_0(x) > 0 & \\ w(x,0) = w_0(x) > 0 & \text{for } x\in\overline{\Omega}, \end{cases} \tag{1.1}$$

where $p(x,t)$ is the particle density of a particular species and $w(x,t)$ is the density of the local control species.

From the boundary condition and the first equation of (1.1), we can easily deduce that, for any solution $p(x,t)$ of (1.1), we have

$$\int_\Omega p(x,t)dx = \int_\Omega p(x,0)dx. \tag{1.2}$$

[*] Received by the editors September 11, 2000; accepted for publication (in revised form) May 14, 2001; published electronically December 7, 2001. This research was supported by the NNSF of China, The Royal Society of London, and the LI Foundation.

http://www.siam.org/journals/sima/33-4/33779.html

[†] School of Mathematics, Wuhan University, Wuhan, 430072, People's Republic of China (ynyang@public.wh.hb.cn, chenhua@whu.edu.cn, liuweian@public.wh.hb.cn).

Through a large number of numerical experiments, Othmer and Stevens found that the asymptotic behavior of the solutions depends strongly on the dynamics of w. In particular, the growth of w determines whether or not blow-up occurs. In fact, from their numerical results in [4], Othmer and Stevens conjecture that, when w has linear growth, there is a global solution of the dynamics, and $p(x,t)$ collapses in some cases, where we say that the solution $p(x,t)$ collapses, which means when t tends to infinity, $p(x,t)$ could be controlled by the initial data, i.e., $\lim_{t\to\infty} \sup \|p(\cdot,t)\|_{L^\infty} < \|p_0(\cdot)\|_{L^\infty}$ (see Definition 2.1 below); whereas when w grows exponentially, $p(x,t)$ should blow up in finite time. For the case of w possessing exponential growth, Levine and Sleeman [2] have studied a special one-dimensional example under the additional boundary-value condition $p_x = w_x = 0$. In that situation they constructed an exact solution supporting the numerical observations of Othmer and Stevens.

In this paper, we extend some results of [2] to the case of general boundary conditions, general positive initial data, and higher dimensional spaces $\mathbb{R}^n$ $(n \geq 1)$. We concentrate on the two situations where w grows linearly and exponentially, respectively. We found that if w grows linearly, the conjectures suggested by numerical observations of Othmer and Stevens are true in general cases. When w possesses exponential growth, the numerical results lose a lot of information. We construct both global and blow-up in a finite-time solution, respectively, for the case Levine and Sleeman [2] considered. So even at the same growth rate the behavior of the biological systems can be very different just because they start their action in different conditions. That may be a very important fact in biological systems.

Our results are the following.

(1) When the production of w is proportional to the local density of p, i.e.,

$$\frac{\partial w}{\partial t} = \beta p - \mu w, \tag{1.3}$$

there exists a unique global solution that is bounded. Furthermore, when the initial data satisfy some additional conditions, in particular, when $w_0(x)$ is a constant, but not necessarily $p_0(x)$, then p collapses.

(2) When the control species grows exponentially, i.e.,

$$\frac{\partial w}{\partial t} = (\beta p - \mu) w, \tag{1.4}$$

both global and blow-up in finite-time solutions may exist dependent on their choice of initial data. We have constructed two families of solutions: global and blow-up in finite time. Also, we have proved that if blow-up occurs, $p(x,t)$, as well as $w(x,t)$, will blow up at the same points.

We discuss linear and exponential growth of w in sections 2–4, respectively. In section 5, we show how our method can be used to solve other dynamics which also arise from mathematical models in biology. We list without proof the main results we have obtained for these new systems.

The main tools we use here are the maximum principle, subsuper solutions, and function transformations.

2. The dynamics with *w* possessing linear growth. In this section, we consider the dynamics with w possessing linear growth. Thus we investigate the following problem:

$$
(2.1)\qquad \begin{cases} \dfrac{\partial p}{\partial t} = D\nabla\cdot\left(p\nabla\left(\ln\left(\dfrac{p}{w}\right)\right)\right) & \\ \dfrac{\partial w}{\partial t} = \beta p - \mu w & \text{for } x\in\Omega,\quad t>0, \\ p\nabla\left(\ln\left(\dfrac{p}{w}\right)\right)\cdot\vec{n} = 0 & \text{for } x\in\partial\Omega,\quad t>0, \\ p(x,0) = p_0(x) > 0 & \\ w(x,0) = w_0(x) > 0 & \text{for } x\in\Omega, \end{cases}
$$

where Ω is a bounded smooth domain in $\mathbb{R}^n$ $(n\geq 1)$, $\vec{n}$ is the outer normal on $\partial\Omega$, and $D>0$, $\beta>0$, $\mu\geq 0$ are all constants.

In their numerical results, Othmer and Stevens [4] suggest that when $\mu>0$, there is a global solution that might collapse. In the case $\mu=0$, the asymptotic behavior of the solution is still unknown, but it is conjectured that small amplitude stable solutions exist.

In order to prove their conjecture, we introduce a new function $u(x,t)=\frac{p(x,t)}{w(x,t)}$. It can be easily shown that $(p(x,t),w(x,t))$ is a solution of the dynamics (2.1) if and only if $(u(x,t),w(x,t))$ is a solution of the following dynamics:

$$
(2.2)\qquad \begin{cases} \dfrac{\partial u}{\partial t} = D\Delta u + D\dfrac{1}{w}(\nabla w)\cdot(\nabla u) + (\mu-\beta u)u & \\ w(x,t) = w_0(x)\exp\left\{\displaystyle\int_0^t(\beta u(x,\tau)-\mu)\,d\tau\right\} & \text{for } x\in\Omega,\quad t>0, \\ \dfrac{\partial u}{\partial n} = 0 & \text{for } x\in\partial\Omega,\quad t>0, \\ u(x,0) = u_0(x) \triangleq \dfrac{p_0(x)}{w_0(x)} > 0 & \\ w(x,0) = w_0(x) > 0 & \text{for} x\in\Omega \end{cases}
$$

and

$$
(2.3)\quad p(x,t) = w_0(x)u(x,t)\exp\left\{\int_0^t(\beta u(x,\tau)-\mu)\,d\tau\right\}\qquad \text{for } x\in\Omega,\quad t>0.
$$

Here we have used the fact that

$$
\begin{aligned} u_t &= \left(\frac{p}{w}\right)_t = \frac{1}{w}(p_t - uw_t), \\ p_t &= D\nabla\cdot\left(p\nabla\left(\ln\left(\frac{p}{w}\right)\right)\right) = D\nabla\cdot(w\nabla u) = Dw\Delta u + D(\nabla w)\cdot(\nabla u). \end{aligned}
$$

Since we can assert the local-in-time existence as well as the uniqueness of solutions by the result of [7] (also see [1, 2, 4]), we set

$$
(2.4)\qquad T = \sup\{\tilde{T}>0; (p,w) \text{ exists for } x\in\Omega, t\in[0,\tilde{T}]\}.
$$

So $T>0$ is well defined. Thus there exists a solution $(u(x,t),w(x,t))$ of the dynamics (2.2) for $x\in\Omega$, $t\in[0,T)$. For fixed w, we investigate the property of u. First we consider $\mu>0$.

LEMMA 2.1. *If $u_0(x)\leq\frac{\mu}{\beta}$, then $u(x,t)\leq\frac{\mu}{\beta}$ as long as the solution exists in time. Furthermore, if for any fixed $t\in(0,T)$, $u(\cdot,t)$ takes its minimum value at the point x_t, then $u_t(x_t,t)>0$ if $u(x_t,t)<\frac{\mu}{\beta}$.*

Proof. Assume that the result is false; then there exists $(x^*, t^*) \in \Omega \times (0, T)$ such that $u(x^*, t^*) > \frac{\mu}{\beta}$. Without loss of generality, we have

$$u(\bar{x}, \bar{t}) = \max_{\bar{\Omega} \times [0, t^*]} u(x, t) > \frac{\mu}{\beta}.$$

We know $\bar{t} > 0$ because of $u(\bar{x}, 0) = u_0(\bar{x}) \leq \frac{\mu}{\beta}$. Since at the point $(x, t) = (\bar{x}, \bar{t})$ we have

$$u_t - D\Delta u - D\frac{1}{w}(\nabla w) \cdot (\nabla u) = (\mu - \beta u)u < 0,$$

we get $(\bar{x}, \bar{t}) \notin \Omega \times (0, t^*]$ by the maximum principle. That means $\bar{x} \in \partial\Omega$. Again, by the maximum principle argument, we should have $\frac{\partial u}{\partial n}\big|_{(\bar{x}, \bar{t})} > 0$, which is a contradiction. That leads to $u(x, t) \leq \frac{\mu}{\beta}$ for all $(x, t) \in \bar{\Omega} \times (0, T)$.

For any fixed $t > 0$ we obtain from the maximum principle

$$-D\Delta u - D\frac{1}{w}(\nabla w) \cdot (\nabla u) = -u_t + (\mu - \beta u)u \leq 0$$

at the point (x_t, t). Since $(\mu - \beta u)u \geq 0$ for all $(x, t) \in \Omega \times (0, T)$, we have $u_t \geq 0$ at (x_t, t). When $u(x_t, t) < \frac{\mu}{\beta}$, we can get $u_t(x_t, t) > 0$ immediately. This completes the proof. □

It is easy to see that similar arguments lead to the following results.

Lemma 2.2. *If $u_0(x) \geq \frac{\mu}{\beta}$, then $u(x, t) \geq \frac{\mu}{\beta}$ for all $(x, t) \in \overline{\Omega} \times [0, T)$. If $u(\cdot, t)$ takes its maximum at the point $x = x^t$ for any fixed $t \in (0, T)$, then $u_t(x^t, t) < 0$ if $u(x^t, t) > \frac{\mu}{\beta}$.*

Lemma 2.3. *If $m \triangleq \min_{x \in \overline{\Omega}} u_0(x) < \frac{\mu}{\beta} < \max_{x \in \overline{\Omega}} u_0(x) \triangleq M$, then $m \leq u(x, t) \leq M$ for all $(x, t) \in \overline{\Omega} \times (0, T)$. Furthermore, $u_t(x^t, t) < 0$ if $u(x^t, t) > \frac{\mu}{\beta}$, and $u_t(x_t, t) > 0$ if $u(x_t, t) < \frac{\mu}{\beta}$, where x^t, x_t are the same as above.*

Remark 2.1. If $T = +\infty$, we can prove that $\lim_{t \to +\infty} u(x, t) = \frac{\mu}{\beta}$ for $\mu > 0$.

Theorem 2.1. *For arbitrary strictly positive initial data $(p_0(x), w_0(x))$, $x \in \overline{\Omega}$, compatible with the boundary condition along $x \in \partial\Omega$, there exists a unique global positive solution $(p(x, t), w(x, t))$ of the dynamics* (2.1) *when $\mu > 0$.*

Proof. We know that the dynamics (2.1) have a unique local positive solution. What we need to do is to prove $T = +\infty$.

Assume $T < +\infty$. From Lemmas 2.1–2.3 we know that the solution $u(x, t)$ of the related dynamics (2.2) is a bounded smooth function in $\Omega \times (0, T)$. So the function $w(x, t) = w_0(x) \exp\{\int_0^t (\beta u(x, \tau) - \mu)\, d\tau\}$ is not only well defined but also continuous on $\bar{\Omega} \times [0, T]$. This implies that both $u(x, t)$ and $p(x, t)$ are well defined and continuous on $\bar{\Omega} \times [0, T]$. Replacing the initial data $(p_0(x), w_0(x))$ of the dynamics (2.1) by $(p(x, T), w(x, T))$, we can obtain a new local positive solution $(\bar{p}(x, t), \bar{w}(x, t))$ defined in $\Omega \times [0, \delta)$ for some $\delta > 0$. It is obvious that $(P(x, t), W(x, t))$ is a positive solution of the dynamics (2.1) defined in $\Omega \times [0, T + \delta)$, where

$$P(x, t) = \begin{cases} p(x, t) & \text{for } (x, t) \in \Omega \times (0, T), \\ \bar{p}(x, t - T) & \text{for } (x, t) \in \Omega \times [T, T + \delta), \end{cases}$$

$$W(x, t) = \begin{cases} w(x, t) & \text{for } (x, t) \in \Omega \times (0, T), \\ \bar{w}(x, t - T) & \text{for } (x, t) \in \Omega \times [T, T + \delta). \end{cases}$$

This result is a contradiction to the definition of T. Thus $T = +\infty$, as required. $\square$

The uniqueness of the solution for (2.1) is trivial.

Remark 2.2. The problem (2.1) is actually a special case of the following Keller–Segel (KS) model with Neumann boundary conditions, i.e.,

$$\text{(KS)}\begin{cases} p_t = \nabla \cdot (d_1 \nabla p - \chi p \nabla \log w), \qquad w_t = d_2 \triangle w + \beta p - \mu w, \\ \text{Neumann boundary conditions,} \end{cases}$$

under the conditions of $d_1 = \chi = D,\ d_2 = 0$. From Theorem 2.1, we have proved, in this case, that the (KS) model has a unique global solution. However, the global existence of solutions for the general (KS) models depends on the parameters in (KS), the space dimensions, and initial functions (cf. [13, 14, 15]).

Based on their numerical results for a wide variety of initial data, Othmer and Stevens conjectured in [4] that $p(x,t)$ collapse when $\mu > 0$. In the case $\mu = 0$, this is unknown. Before discussing this problem, we introduce the concept of collapse used in [4].

DEFINITION 2.1. *Let $p(x,t)$ be the solution of the dynamics* (2.1) *for given initial distribution $p_0(x)$. Then if $\lim_{t\to\infty} \sup \|p(\cdot,t)\|_{L^\infty} < \|p_0(\cdot)\|_{L^\infty}$, we say there is collapse.*

In other words, if $p(x,t)$ denotes the particle density of a particular species and $p(x,t)$ collapses, what represents the species will not be aggregation (see [4] for details).

It is obvious that if $\mu > 0$, the dynamics (2.1) has a constant solution $(p, w) = (p_0, \beta p_0/\mu)$. So collapse does not occur. Furthermore, we can give a class of steady-state solutions for p which do not collapse at all.

Example 2.1. For any smooth function $\Phi(x) > 0$ defined on $\overline{\Omega}$,

$$(p, w) = \left(1, e^{-\mu t}\left(c - \frac{\beta}{\mu}\right) + \frac{\beta}{\mu}\right)\Phi(x)$$

is the solution of the dynamics (2.1) with initial data $(p_0(x), w_0(x)) = (1, c)\Phi(x)$ for all $c > 0$. It is obvious that we cannot expect collapse.

THEOREM 2.2. *Let $\mu > 0$. For the dynamics* (2.1) *the solution (p, w) is bounded. Furthermore, for all $(x,t) \in \overline{\Omega} \times [0, +\infty)$ we have $p(x,t) < w_0(x) \max_{\overline{\Omega}} u_0(x)$ if $w_0(x) \neq cp_0(x)$ for some $c > 0$.*

Proof. Let $c_0 = [\max_{x\in\bar{\Omega}} u_0(x)]^{-1}$ and $v(t) = e^{\mu t}[c_0 + \frac{\beta}{\mu}(e^{\mu t} - 1)]^{-1}$. We can easily find that $v(t)$ is a supersolution of the dynamics (2.2). From $u(x,t) \le v(t)$ for all $(x,t) \in \Omega \times (0 + \infty)$ we have

$$\begin{aligned} p(x,t) &= w_0(x) u(x,t) \exp\left\{\int_0^t (\beta u(x,\tau) - \mu)\, d\tau\right\} \\ &\le w_0(x) v(t) \exp\left\{\int_0^t (\beta v(\tau) - \mu)\, d\tau\right\} \\ &= w_0(x)/c_0 \\ &= w_0(x) \max_{x\in\bar{\Omega}} u_0(x) < +\infty. \end{aligned}$$

Next, from $(e^{\mu t} w)_t = \beta e^{\mu t} p$, we have

$$\begin{aligned} e^{\mu t}w(x,t) &= w(x,0) + \beta \int_0^t e^{\mu\tau} p(x,\tau)d\tau \\ &\leq w(x,0) + \beta \int_0^t w_0(x) \max_{x\in\bar{\Omega}} u_0(x) e^{\mu\tau} d\tau \\ &= w_0(x) + \frac{\beta}{\mu} w_0(x) \max_{x\in\bar{\Omega}} u_0(x)[e^{\mu t} - 1], \end{aligned}$$

which means

$$w(x,t) \leq e^{-\mu t}w_0(x) + \frac{\beta}{\mu} w_0(x) \max_{x\in\bar{\Omega}} u_0(x)[1 - e^{-\mu t}] < +\infty.$$

When $w_0(x) \neq cp_0(x)$, then $u(x,t) < v(t)$ for all $(x,t) \in \Omega \times (0,+\infty)$, which implies $p(x,t) < w_0(x) \max_{x\in\bar{\Omega}} u_0(x)$ for $(x,t) \in \Omega \times (0,+\infty)$. This completes the proof. □

Let $u_0(x)$ attain its maximum at the point x_1 and $p_0(x)$ its maximum at the point x_2. Denote $\sigma = p_0(x_2)/p_0(x_1)$. If $w_0(x) \leq \sigma w_0(x_1)$ for all $x \in \bar{\Omega}$, then $p(x,t) < p_0(x_2)$ when $w_0(x) \neq cp_0(x)$. This fact leads to the following result on the collapse.

COROLLARY 2.1. *Let $\mu > 0$, and suppose that $u_0(x)$ takes its maximum at the point x_1, and let $\sigma = [p_0(x_1)]^{-1} \max_{x\in\overline{\Omega}} p_0(x)$. If $w_0(x) \leq \sigma w_0(x_1)$ for $x \in \overline{\Omega}$ when $w_0(x) \neq cp_0(x)$ for some $c > 0$, then $p(x,t) < \max_{x\in\overline{\Omega}} p_0(x)$ for all $(x,t) \in \overline{\Omega} \times (0,+\infty)$. This means that there is collapse.*

Othmer and Stevens discuss in [4] the system

$$\begin{cases} \dfrac{\partial p}{\partial t} = D\dfrac{\partial}{\partial x}\left(p\dfrac{\partial}{\partial x}\left(\ln\left(\dfrac{p}{w}\right)\right)\right) & \text{for } x \in (0,1), \quad t > 0, \\ \dfrac{\partial w}{\partial t} = p - \mu w & \\ p\dfrac{\partial}{\partial x}\left(\ln\left(\dfrac{p}{w}\right)\right) = 0 & \text{for } x = 0, 1, \quad t > 0, \\ p(x,0) = p_0(x) > 0 & \text{for } x \in [0,1] \\ w(x,0) = w_0(x) > 0 & \end{cases}$$

and find that if $\mu = 0$, the space-independent solution $(p_0, w_0 + p_0 t)$ is unstable.

When $\mu > 0$, we have seen that there exists a solution

$$(p,w) = \left(1, e^{-\mu t}\left(c - \frac{\beta}{\mu}\right) + \frac{\beta}{\mu}\right)\Phi(x) \tag{2.5}$$

for the dynamics (2.1), where p is t-independent, with initial data $(p_0, w_0) = (1, c)\Phi(x)$. However, we can prove the solution (2.5) will be unstable asymptotically. In fact, if we choose a positive function $\psi(x)$, which takes its minimum at the maximum point of the function Φ, then both functions Φ and $\frac{\Phi}{1+\psi}$ attain their maxima at the same point. Thus, as a consequence of Theorems 2.1 and 2.2, we know that the dynamics (2.1) has a unique global solution $(p(x,t), w(x,t))$ with initial data $(p_0, w_0) = (\frac{1}{1+\psi(x)}, c)\Phi(x)$, which satisfies $\frac{p(x,t)}{w(x,t)} = u(x,t) < v(t)$ and

$$\begin{aligned} p(x,t) &< w_0(x) \max_{x\in\bar{\Omega}} u_0(x) = w_0(x) \max_{x\in\bar{\Omega}} \left(\frac{1}{c(1+\psi(x))} \right) \\ &= c\Phi(x) \max_{x\in\bar{\Omega}} \left(\frac{1}{c(1+\psi(x))} \right) \le \max_{x\in\bar{\Omega}} \Phi(x) \max_{x\in\bar{\Omega}} \left(\frac{1}{1+\psi(x)} \right) \\ &= \max_{x\in\bar{\Omega}} \left(\frac{\Phi(x)}{1+\psi(x)} \right). \end{aligned}$$

Let $\bar{x}$ be the maximum point of $p_0(x)$ (i.e., $\bar{x}$ is the maximum point of $\Phi(x)$ and the minimum point of $\psi(x)$); then we have

$$p(\bar{x},t) < p_0(\bar{x}) = \frac{\Phi(\bar{x})}{1+\psi(\bar{x})} < \Phi(\bar{x}).$$

That implies, for any $\psi(x) > 0$ but quite small, we have

$$\overline{\lim_{t\to+\infty}}\, p(\bar{x},t) \le p_0(\bar{x}) < \Phi(\bar{x}).$$

Therefore, we get the following result.

COROLLARY 2.2. *When $\mu > 0$, the solution*

$$(p,w) = \left(1, e^{-\mu t}\left(c - \frac{\beta}{\mu}\right) + \frac{\beta}{\mu}\right)\Phi(x)$$

with t-independent p is not asymptotically stable.

Next we consider the case $\mu = 0$.

THEOREM 2.3. *When $\mu = 0$, for any strictly positive initial data $(p_0(x), w_0(x))$ (satisfying the comparison condition similarly as above in Theorem* 2.1*), there exists a unique global positive solution $(p(x,t), w(x,t))$ of the dynamics* (2.1) *such that $p(x,t)$ is bounded and $w(x,t) \overset{t\to+\infty}{\longrightarrow} +\infty$. Furthermore,*

$$p(x,t) < w_0(x) \max_{\overline{\Omega}} u_0(x)$$

if $w_0(x) \neq cp_0(x)$ for some $c > 0$.

Proof. Let $u(x,t) = \frac{p(x,t)}{w(x,t)}$; we know that $(p(x,t), w(x,t))$ is a positive solution of the dynamics (2.1) if and only if $(u(x,t), w(x,t))$ is a positive solution of the following dynamics:

$$\begin{cases} \left.\begin{aligned} &\frac{\partial u}{\partial t} = D\Delta u + D\frac{1}{w}(\nabla w)\cdot(\nabla u) - \beta u^2 \\ &w(x,t) = w_0(x)\exp\left\{\int_0^t (\beta u(x,\tau))\,d\tau\right\} \end{aligned}\right\} & \text{for } x\in\Omega, \quad t>0, \\ \dfrac{\partial u}{\partial n} = 0 & \text{for } x\in\partial\Omega, \quad t>0, \\ \left.\begin{aligned} &u(x,0) = u_0(x) \triangleq \frac{p_0(x)}{w_0(x)} > 0 \\ &w(x,0) = w_0(x) > 0 \end{aligned}\right. & \text{for } x\in\overline{\Omega} \end{cases}$$

and

$$p(x,t) = w_0(x)u(x,t)\exp\left\{\int_0^t \beta u(x,\tau)\,d\tau\right\}$$

for $x \in \Omega$, $t > 0$. It can be easily found that

$$\left(\min_{x\in\bar\Omega} u_0(x)\right)\left[1+\beta \min_{x\in\bar\Omega} u_0(x)t\right]^{-1} \le u(x,t) \le \left(\max_{x\in\bar\Omega} u_0(x)\right)\left[1+\beta \max_{x\in\bar\Omega} u_0(x)t\right]^{-1},$$

which implies that there exists a global positive solution for the dynamics (2.1) by the method we use in the proof of Theorem 2.1. If $w_0(x) \ne cp_0(x)$ for some $c > 0$, we obtain

$$u(x,t) < \left(\max_{x\in\bar\Omega} u_0(x)\right)\left[1+\beta \max_{x\in\bar\Omega} u_0(x)t\right]^{-1}$$

and

$$\begin{aligned} p(x,t) &= w_0(x)u(x,t)\exp\left\{\int_0^t \beta u(x,\tau)\,d\tau\right\} \\ &< w_0(x)\frac{\max_{x\in\bar\Omega} u_0(x)}{1+\beta\max_{x\in\bar\Omega} u_0(x)t}\exp\left\{\int_0^t \frac{\beta\max_{x\in\bar\Omega} u_0(x)}{1+\beta\max_{x\in\bar\Omega} u_0(x)\tau}\,d\tau\right\} \\ &= w_0(x)\max_{x\in\bar\Omega} u_0(x), \end{aligned}$$

and $w(x,t) \ge w_0(x)[1+\beta\min_{x\in\bar\Omega} u_0(x)t] \overset{t\to+\infty}{\longrightarrow} +\infty$.

Similarly, we have the following result.

COROLLARY 2.3. *For $\mu = 0$, the statements of Corollaries* 2.1 *and* 2.2 *are also valid.*

3. One-dimensional case with w possessing exponential growth. The growth rate of w is very important to the characteristics of p, which determines whether or not blow-up occurs. Actually, Othmer and Stevens [4] conjectured that, when w grows exponentially, $p(x,t)$ should blow up in finite time. It is obvious that when both initial data $p(x,0)$ and $w(x,0)$ are positive constants, we can get a global solution immediately: $(p_0, w_0e^{(\beta p_0-\mu)t})$. What we are very interested in is the situation about nonconstant initial data. In a recent result of [2], Levine and Sleeman studied a special one-dimensional example under the additional boundary-value condition $p_x = w_x = 0$, and they constructed a class of solutions

$$\begin{aligned} p(x,t) &= 1 - 2Nc\varepsilon e^{Nct}\frac{\varepsilon e^{Nct}-\cos(Nx)}{1-2\varepsilon e^{Nct}\cos(Nx)+\varepsilon^2e^{2Nct}}, \\ w(x,t) &= \frac{e^t}{1-2\varepsilon e^{Nct}\cos(Nx)+\varepsilon^2e^{2Nct}} \end{aligned}$$

satisfying the problem, taking $D = \beta = 1, \mu = 0$,

$$\begin{cases} \dfrac{\partial p}{\partial t} = D\dfrac{\partial}{\partial x}\left(p\dfrac{\partial}{\partial x}\left(\ln\left(\dfrac{p}{w}\right)\right)\right) & \\ & \text{for } x \in (0,\pi), \quad t > 0, \\ \dfrac{\partial w}{\partial t} = (\beta p - \mu)w & \\ p_x = w_x = 0 & \text{for } x = 0, \pi; \quad t > 0, \\ p(x,0) = \dfrac{1-2\varepsilon(1-Nc)\cos(Nx)+(1-2Nc)\varepsilon^2}{1-2\varepsilon\cos(Nx)+\varepsilon^2} & \\ & \text{for } x \in [0,\pi], \\ w(x,0) = \dfrac{1}{1-2\varepsilon\cos(Nx)+\varepsilon^2} & \end{cases}$$

where $N > 0$ is an integer, $c = \frac{2}{N+\sqrt{N^2+4}}$, and $\varepsilon > 0$ sufficiently small. Levine and Sleeman's work confirms the conjecture of Othmer and Stevens. Unfortunately, combining the technique of [2] and the other technique, even we can find that, in the one-dimensional case, the blow-up conjecture is false. In fact, the situation in this case is more complicated; we can find two kinds of solutions. One would exist globally, and another one would blow up in finite time. In particular, the solution as obtained by Levine and Sleeman [2] is one of the special cases in our blow-up solutions family here.

At the beginning, let us follow the technique in Levine and Sleeman [2] and consider the special case of the dynamics:

$$(3.1)\qquad \begin{cases} \dfrac{\partial p}{\partial t} = \dfrac{\partial}{\partial x}\left(p\dfrac{\partial}{\partial x}\left(\ln\left(\dfrac{p}{w}\right)\right)\right) & \text{for } x \in (0,\pi), \quad t > 0, \\ \dfrac{\partial w}{\partial t} = pw & \\ p_x = w_x = 0 & \text{for } x = 0, \pi; \quad t > 0, \\ p(x,0) = p_0(x) > 0 & \\ w(x,0) = w_0(x) > 0 & \text{for } x \in [0,\pi]. \end{cases}$$

Let $\psi(x,t) = \ln w(x,t)$; we obtain the following initial boundary-value problem for ψ:

$$(3.2)\qquad \begin{cases} \psi_{tt} - \psi_{xxt} + (\psi_x\psi_t)_x = 0 & \text{for } x \in (0,\pi), \quad t > 0, \\ \psi_x = 0 & \text{for } x = 0, \pi; \quad t > 0, \\ \psi(x,0) = \psi_0(x) = \ln w_0(x) & \\ \psi_t(x,0) = p_0(x) > 0 & \text{for } x \in (0,\pi). \end{cases}$$

Let $\psi = \alpha t + \phi$. Then ϕ satisfies

$$(3.3)\qquad \begin{cases} \phi_{tt} + \alpha\phi_{xx} - \phi_{xxt} + \phi_t\phi_{xx} + \phi_x\phi_{xt} = 0 & \text{for } 0 < x < \pi, \quad t > 0, \\ \phi_x = 0 & \text{for } x = 0, \pi; \quad t > 0, \\ \phi(x,0) = \phi_0(x) & \\ \phi_t(x,0) = p_0(x) - \alpha & \text{for } x \in (0,\pi). \end{cases}$$

Let $\phi(x,t) = -\ln(c-u)$. Then

$$\phi_t = \frac{u_t}{c-u}, \quad \phi_x = \frac{u_x}{c-u}, \quad \phi_{tt} = \frac{(c-u)u_{tt} + u_t^2}{(c-u)^2},$$

$$\phi_{xx} = \frac{(c-u)u_{xx} + u_x^2}{(c-u)^2}, \quad \phi_{xt} = \frac{(c-u)u_{xt} + u_xu_t}{(c-u)^2},$$

and

$$\begin{aligned} \phi_{xxt} &= \frac{1}{(c-u)^4}\big\{(c-u)^2\big[(c-u)u_{xxt} - u_tu_{xx} + 2u_xu_{xt}\big] \\ &\quad +2(c-u)u_t\big[(c-u)u_{xx} + u_x^2\big]\big\} \\ &= \frac{1}{(c-u)}u_{xxt} + \frac{1}{(c-u)^2}\big[2u_xu_{xt} + u_tu_{xx}\big] + \frac{2u_tu_x^2}{(c-u)^3}. \end{aligned}$$

So we have

$$\begin{aligned}0 &= \phi_{tt} + \alpha\phi_{xx} - \phi_{xxt} + \phi_t\phi_{xx} + \phi_x\phi_{xt}\\ &= \frac{1}{(c-u)^2}\big(u_t^2 + \alpha u_x^2 - u_x u_{xt}\big) + \frac{1}{c-u}\big(u_{tt} + \alpha u_{xx} - u_{xxt}\big)\\ &= \frac{1}{(c-u)^2}\big(u_t^2 + \alpha u_x^2 - u_x u_{xt} + cu_{tt} + c\alpha u_{xx}\\ &\quad -cu_{xxt} - uu_{tt} - \alpha uu_{xx} + uu_{xxt}\big).\end{aligned}$$

If $u(x,t) < c$, the above equation is equivalent to the following equation:

$$cu_{tt} + c\alpha u_{xx} - cu_{xxt} + u_t^2 + \alpha u_x^2 - u_x u_{xt} - uu_{tt} - \alpha uu_{xx} + uu_{xxt} = 0.$$

Thus we have

$$\text{(3.4)}\quad \begin{cases} cu_{tt} + c\alpha u_{xx} - cu_{xxt} = & -\big(u_t^2 + \alpha u_x^2 - u_x u_{xt} - uu_{tt} - \alpha uu_{xx} + uu_{xxt}\big)\\ & \text{for } 0 < x < \pi, \quad t > 0,\\ u_x = 0 & \text{for } x = 0, \pi, \quad t > 0.\end{cases}$$

Next, let $u(x,t) = X(x)T(t) + g(t)$. Then

$$\begin{aligned}0 &= cXT'' + cg''(t) + c\alpha X''T - cX''T' + \big[XT' + g'(t)\big]^2 + \alpha X'^2T^2 - X'^2TT'\\ &\quad -\big(XT + g(t)\big)\big(XT'' + g''(t)\big) - \alpha\big[XT + g(t)\big]X''T + \big[XT + g(t)\big]X''T'\\ &= \big[cT'' + 2g'(t)T' - g''(t)T - g(t)T''\big]X + \big[c\alpha T - cT' - \alpha g(t)T + g(t)T'\big]X''\\ &\quad +\big[T'^2 - TT''\big]X^2 + \big[\alpha T^2 - TT'\big]X'^2 + \big[TT' - \alpha T^2\big]XX''\\ &\quad +[g'(t)]^2 - g(t)g''(t) + cg''(t).\end{aligned}$$

If $T'^2 = TT''$, we obtain $\frac{T'}{T} = \frac{T''}{T'}$, which implies $T(t) = c_2e^{c_1t}$. In this case, we can get

$$\begin{aligned}0 &= \big\{\big[cc_1^2 + 2c_1g'(t) - g''(t) - c_1^2g(t)\big]X + \big[\alpha c - cc_1 - \alpha g(t) + c_1g(t)\big]X''\\ &\quad +\big[\alpha - c_1\big](X')^2T + \big[c_1 - \alpha\big]TXX''\big\}T + cg''(t) - g(t)g''(t) + [g'(t)]^2.\end{aligned}$$

Choose $X(x) = \cos nx$. Then

$$\begin{aligned}0 &= \big[cc_1^2 + 2c_1g'(t) - g''(t) - c_1^2g(t) - \alpha cn^2 + cc_1n^2 + \alpha n^2g(t) - c_1n^2g(t)\big]XT\\ &\quad +\big\{n^2[\alpha - c_1]\sin^2 nx + n^2[\alpha - c_1]\cos^2 nx\big\}T^2 + cg''(t) - g(t)g''(t) + [g'(t)]^2\\ &= \big[cc_1^2 - \alpha cn^2 + cc_1n^2 + 2c_1g'(t) - g''(t) + (\alpha n^2 - c_1^2 - c_1n^2)g(t)\big]XT\\ &\quad +(\alpha n^2 - c_1n^2)T^2 + cg''(t) - g(t)g''(t) + [g'(t)]^2.\end{aligned}$$

Now let us consider the following equations:

$$\begin{cases} g''(t) - 2c_1g'(t) + (c_1^2 - \alpha n^2 + c_1n^2)g(t) = c(c_1^2 + c_1n^2 - \alpha n^2), & \text{(3.5)}\\ (c - g(t))g''(t) + (g'(t))^2 = n^2c_2^2(c_1 - \alpha)e^{2c_1t}. & \text{(3.6)}\end{cases}$$

From (3.5), it is obvious that $g(t) = Ae^{k_1 t} + Be^{k_2 t} + c$, where $c_1 < \alpha$, $k_1 = c_1 + n\sqrt{\alpha - c_1}$, and $k_2 = c_1 - n\sqrt{\alpha - c_1}$. Taking $g(t)$ into (3.6), we have

$$\begin{aligned} n^2 c_2^2 (c_1 - \alpha) e^{2c_1 t} &= \big[- Ae^{k_1 t} - Be^{k_2 t} \big] \big[Ak_1^2 e^{k_1 t} + Bk_2^2 e^{k_2 t} \big] + \big[Ak_1 e^{k_1 t} + Bk_2 e^{k_2} \big]^2 \\ &= -AB(k_1 - k_2)^2 e^{2c_1 t}, \end{aligned}$$

which implies

$$n^2 c_2^2 (\alpha - c_1) = AB(k_1 - k_2)^2 = 4n^2 AB(\alpha - c_1),$$

which leads to $4AB = c_2^2$. Replacing $2c_2$ by c_2, A by $-A$, and B by $-B$, we obtain

$$\psi(x,t) = \alpha t - \ln[\pm 2\sqrt{AB} e^{c_1 t} \cos nx + Ae^{k_1 t} + Be^{k_2 t}]$$

and

$$p(x,t) = \alpha - \frac{Ak_1 e^{k_1 t} + Bk_2 e^{k_2 t} \pm 2c_1 \sqrt{AB} e^{c_1 t} \cos nx}{Ae^{k_1 t} + Be^{k_2 t} \pm 2\sqrt{AB} e^{c_1 t} \cos nx}.$$

It is obvious that

$$w(x,t) = \frac{e^{\alpha t}}{\pm 2\sqrt{AB} e^{c_1 t} \cos nx + Ae^{k_1 t} + Be^{k_2 t}}.$$

(p, w) is the solution for the problem (3.1) with the initial data

$$(p_0(x), w_0(x)) = \left(\alpha - \frac{Ak_1 + Bk_2 \pm 2c_1 \sqrt{AB} \cos nx}{A + B \pm 2\sqrt{AB} \cos nx}, \frac{1}{\pm 2\sqrt{AB} \cos nx + A + B} \right).$$

THEOREM 3.1. *For the dynamics* (3.1), *we have the following kinds of solutions:*

(a) *Assume $A > B > 0$; if $c_1 < \alpha - n^2 \{ \frac{\sqrt{A}+\sqrt{B}}{\sqrt{A}-\sqrt{B}} \}^2$, then $(p(x,t), w(x,t))$ is the global solution for the dynamics* (3.1).

(b) *Assume $0 < A < B$; then for any $c_1 < \alpha$ there exists $T > 0$ such that $(p(x,t), w(x,t))$ is the solution of the dynamics* (3.1) *for $0 < t < T$, and the solution will blow up at the finite time $t = T$ at some point $x \in [0, \pi]$.*

Proof. (a) Since $A > B > 0$, we can choose a point x_0, satisfying $\cos nx_0 = -1$. Then, for $x \in [0, \pi], t > 0$, we have

$$\begin{aligned} Ae^{k_1 t} + Be^{k_2 t} \pm 2\sqrt{AB} e^{c_1 t} \cos nx &\geq Ae^{k_1 t} + Be^{k_2 t} + 2\sqrt{AB} e^{c_1 t} \cos nx_0 \\ &= e^{c_1 t} \left(\sqrt{A} e^{\frac{(n\sqrt{\alpha - c_1})t}{2}} - \sqrt{B} e^{-\frac{(n\sqrt{\alpha - c_1})t}{2}} \right)^2 \\ &\geq e^{c_1 t} (\sqrt{A} - \sqrt{B})^2 > 0, \end{aligned}$$

which implies that $p(x,t)$ and $w(x,t)$ are well defined and $w(x,t) > 0$ for all $t > 0$. Furthermore, because of $c_1 < \alpha$ and $c_1 < \alpha - n^2 \{ \frac{\sqrt{A}+\sqrt{B}}{\sqrt{A}-\sqrt{B}} \}^2$, which implies that

$\alpha - c_1 - n\sqrt{\alpha - c_1}\frac{\sqrt{A}+\sqrt{B}}{\sqrt{A}-\sqrt{B}} > 0$, we obtain that

$$
\begin{aligned}
p(x,t) &= \alpha - \frac{Ak_1e^{k_1t} + Bk_2e^{k_2t} \pm 2\sqrt{AB}c_1e^{c_1t}\cos nx}{Ae^{k_1t} + Be^{k_2t} \pm 2\sqrt{AB}e^{c_1t}\cos nx} \\
&= \alpha - \frac{Ak_1e^{n\sqrt{\alpha-c_1}t} + Bk_2e^{-n\sqrt{\alpha-c_1}t} \pm 2\sqrt{AB}c_1\cos nx}{Ae^{n\sqrt{\alpha-c_1}t} + Be^{-n\sqrt{\alpha-c_1}t} \pm 2\sqrt{AB}\cos nx} \\
&= \alpha - c_1 - n\sqrt{\alpha - c_1}\frac{Ae^{n\sqrt{\alpha-c_1}t} - Be^{-n\sqrt{\alpha-c_1}t}}{Ae^{n\sqrt{\alpha-c_1}t} + Be^{-n\sqrt{\alpha-c_1}t} \pm 2\sqrt{AB}\cos nx} \\
&\geq \alpha - c_1 - n\sqrt{\alpha - c_1}\frac{Ae^{n\sqrt{\alpha-c_1}t} - Be^{-n\sqrt{\alpha-c_1}t}}{\left\{\sqrt{A}e^{\frac{n\sqrt{\alpha-c_1}t}{2}} - \sqrt{B}e^{-\frac{n\sqrt{\alpha-c_1}t}{2}}\right\}^2} \\
&= \alpha - c_1 - n\sqrt{\alpha - c_1}\frac{\sqrt{A}e^{\frac{n\sqrt{\alpha-c_1}t}{2}} + \sqrt{B}e^{\frac{-n\sqrt{\alpha-c_1}t}{2}}}{\sqrt{A}e^{\frac{n\sqrt{\alpha-c_1}t}{2}} - \sqrt{B}e^{-\frac{n\sqrt{\alpha-c_1}t}{2}}} \\
&\geq \alpha - c_1 - n\sqrt{\alpha - c_1}\frac{\sqrt{A}+\sqrt{B}}{\sqrt{A}-\sqrt{B}} > 0.
\end{aligned}
$$

Furthermore, for any $x \in [0,\pi], t > 0$, we get that

$$
\begin{aligned}
p(x,t) &= \alpha - c_1 - n\sqrt{\alpha - c_1}\frac{Ae^{n\sqrt{\alpha-c_1}t} - Be^{-n\sqrt{\alpha-c_1}t}}{Ae^{n\sqrt{\alpha-c_1}t} + Be^{-n\sqrt{\alpha-c_1}t} \pm 2\sqrt{AB}\cos nx} \\
&\leq \alpha - c_1 - n\sqrt{\alpha - c_1}\frac{Ae^{n\sqrt{\alpha-c_1}t} - Be^{-n\sqrt{\alpha-c_1}t}}{\left\{\sqrt{A}e^{\frac{n\sqrt{\alpha-c_1}t}{2}} + \sqrt{B}e^{-\frac{n\sqrt{\alpha-c_1}t}{2}}\right\}^2} \\
&= \alpha - c_1 - n\sqrt{\alpha - c_1}\frac{\sqrt{A}e^{\frac{n\sqrt{\alpha-c_1}t}{2}} - \sqrt{B}e^{\frac{-n\sqrt{\alpha-c_1}t}{2}}}{\sqrt{A}e^{\frac{n\sqrt{\alpha-c_1}t}{2}} + \sqrt{B}e^{-\frac{n\sqrt{\alpha-c_1}t}{2}}} \\
&\leq \alpha - c_1 - n\sqrt{\alpha - c_1} < +\infty, \\
w(x,t) &= \frac{e^{\alpha t}}{\pm 2\sqrt{AB}e^{c_1t}\cos nx + Ae^{k_1t} + B^{k_2t}} \\
&\leq \frac{e^{(\alpha-c_1)t}}{-2\sqrt{AB} + Ae^{n\sqrt{\alpha-c_1}t} + Be^{-n\sqrt{\alpha-c_1}t}} \\
&= \frac{e^{(\alpha-c_1)t}}{\left\{\sqrt{A}e^{\frac{n\sqrt{\alpha-c_1}t}{2}} - \sqrt{B}e^{-\frac{n\sqrt{\alpha-c_1}t}{2}}\right\}^2} < +\infty.
\end{aligned}
$$

That leads to the global existence of the solution for the system (3.1).

(b) When $B > A > 0$, there exists a point $x_0 \in [0,\pi]$ such that $\cos nx_0 = \mp 1$

satisfying, for any $0 < t < T = \frac{1}{2n\sqrt{\alpha - c_1}} \ln \frac{B}{A}$,

$$\begin{aligned} p(x_0, t) &= \alpha - c_1 + n\sqrt{\alpha - c_1} \frac{Be^{-n\sqrt{\alpha - c_1}t} - Ae^{n\sqrt{\alpha - c_1}t}}{Be^{-n\sqrt{\alpha - c_1}t} + Ae^{n\sqrt{\alpha - c_1}t} \pm 2\sqrt{AB}\cos nx_0} \\ &= \alpha - c_1 + n\sqrt{\alpha - c_1} \frac{\sqrt{B}e^{\frac{-n\sqrt{\alpha - c_1}t}{2}} + \sqrt{A}e^{\frac{n\sqrt{\alpha - c_1}t}{2}}}{\sqrt{B}e^{\frac{-n\sqrt{\alpha - c_1}t}{2}} - \sqrt{A}e^{\frac{n\sqrt{\alpha - c_1}t}{2}}}, \\ w(x_0, t) &= \frac{e^{\alpha t}}{\pm 2\sqrt{AB}e^{c_1 t}\cos nx_0 + Ae^{k_1 t} + B^{k_2 t}} \\ &= \frac{e^{(\alpha - c_1)t}}{\left\{\sqrt{B}e^{\frac{-n\sqrt{\alpha - c_1}t}{2}} - \sqrt{A}e^{\frac{n\sqrt{\alpha - c_1}t}{2}}\right\}^2}. \end{aligned}$$

It is clear that $\lim_{t \to T^-} p(x_0, t) = +\infty, \lim_{t \to T^-} w(x_0, t) = +\infty$. We can also show that for any $x \in (0, \pi), 0 < t < T$,

$$\begin{aligned} p(x, t) &= \alpha - c_1 + n\sqrt{\alpha - c_1} \frac{Be^{-n\sqrt{\alpha - c_1}t} - Ae^{n\sqrt{\alpha - c_1}t}}{Be^{-n\sqrt{\alpha - c_1}t} + Ae^{n\sqrt{\alpha - c_1}t} \pm 2\sqrt{AB}\cos nx} \\ &\geq \alpha - c_1 + n\sqrt{\alpha - c_1} \frac{\sqrt{B}e^{\frac{-n\sqrt{\alpha - c_1}t}{2}} - \sqrt{A}e^{\frac{n\sqrt{\alpha - c_1}t}{2}}}{\sqrt{B}e^{\frac{-n\sqrt{\alpha - c_1}t}{2}} + \sqrt{A}e^{\frac{n\sqrt{\alpha - c_1}t}{2}}} > 0, \\ w(x, t) &= \frac{e^{\alpha t}}{\pm 2\sqrt{AB}e^{c_1 t}\cos nx + Ae^{k_1 t} + B^{k_2 t}} \\ &\geq \frac{e^{(\alpha - c_1)t}}{\left\{\sqrt{B}e^{\frac{-n\sqrt{\alpha - c_1}t}{2}} + \sqrt{A}e^{\frac{n\sqrt{\alpha - c_1}t}{2}}\right\}^2} > 0, \end{aligned}$$

and

$$\begin{aligned} p(x, t) &= \alpha - c_1 + n\sqrt{\alpha - c_1} \frac{Be^{-n\sqrt{\alpha - c_1}t} - Ae^{n\sqrt{\alpha - c_1}t}}{Be^{-n\sqrt{\alpha - c_1}t} + Ae^{n\sqrt{\alpha - c_1}t} \pm 2\sqrt{AB}\cos nx} \\ &\leq \alpha - c_1 + n\sqrt{\alpha - c_1} \frac{\sqrt{B}e^{\frac{-n\sqrt{\alpha - c_1}t}{2}} + \sqrt{A}e^{\frac{n\sqrt{\alpha - c_1}t}{2}}}{\sqrt{B}e^{\frac{-n\sqrt{\alpha - c_1}t}{2}} - \sqrt{A}e^{\frac{n\sqrt{\alpha - c_1}t}{2}}} < +\infty, \\ w(x, t) &= \frac{e^{\alpha t}}{\pm 2\sqrt{AB}e^{c_1 t}\cos nx + Ae^{k_1 t} + B^{k_2 t}} \\ &\leq \frac{e^{(\alpha - c_1)t}}{\left\{\sqrt{B}e^{\frac{-n\sqrt{\alpha - c_1}t}{2}} - \sqrt{A}e^{\frac{n\sqrt{\alpha - c_1}t}{2}}\right\}^2} < +\infty. \end{aligned}$$

That implies that the solution $(p(x, t), w(x, t))$ of the problem is well defined for $x \in (0, \pi), 0 < t < T$, the solution will blow up at the point $x = x_0$, and t tends to finite-time T.

Theorem 3.1 is proved. □

Remark 3.1. For the different cases corresponding to the different situations in Theorem 3.1, please see Figures 1–4 below.

Remark 3.2. We cannot respect the case $A = B$, because at this time the initial data $(p(x, 0), w(x, 0))$ cannot be well defined at the points x satisfying $\cos nx = \mp 1$.

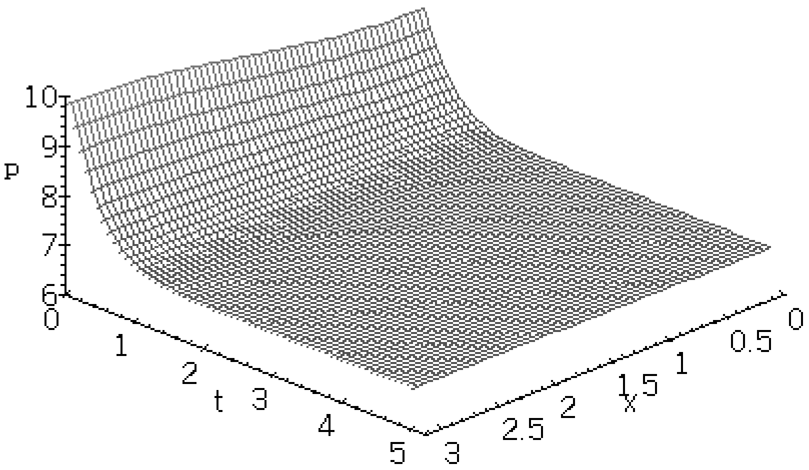


FIG. 1. *In the case of Theorem* 3.1(a)*, for the solution* $p(x,t)$*, where* $A = 4$*,* $B = 1$*,* $n = 1$*,* $c_1 = 1$*,* $\alpha = 11$.

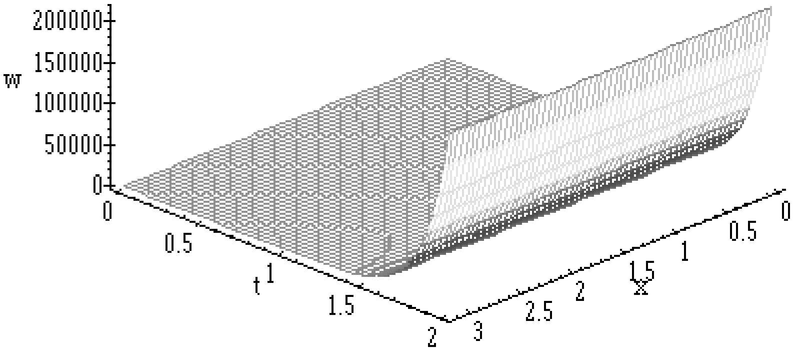


FIG. 2. *In the case of Theorem* 3.1(a)*, for the solution* $w(x,t)$*, where* $A = 4$*,* $B = 1$*,* $n = 1$*,* $c_1 = 1$*,* $\alpha = 11$.

Remark 3.3. If we choose $c_1 = \alpha$, we can obtain $g(t) = (A + Bt)e^t$, and B must be 0 in this case. Then we get $(p(x,t), w(x,t)) = (0, \frac{1}{A+c_2 \cos nx})$ for any $|A| > |c_2|$, which is not permitted in our problem because we want only a positive solution.

Remark 3.4. Assume that $c_1 > \alpha$; we can get the solution

$$\begin{cases} p(x,t) & = \alpha - c_1 + \dfrac{\sqrt{c_1-\alpha}[A \sin n\sqrt{c_1-\alpha}t - B \cos n\sqrt{c_1-\alpha}t]}{A \cos n\sqrt{c_1-\alpha}t + B \sin n\sqrt{c_1-\alpha}t \pm \sqrt{A^2+B^2}\cos nx}, \\ w(x,t) & = \dfrac{e^{(\alpha-c_1)t}}{A \cos n\sqrt{c_1-\alpha}t + B \sin\ n\sqrt{c_1-\alpha}t \pm \sqrt{A^2+B^2}\cos\ nx}. \end{cases}$$

It can be easily found that the initial data $(p(x,0), w(x,0))$ cannot be well defined.

Remark 3.5. Let $c_1 = \frac{-n^2+n\sqrt{n^2+4}}{2}$. Then $k_1 = 2c_1, k_2 = 0$, and

$$p(x,t) = \alpha - 2c_1 \frac{Ae^{2c_1t} + 2\sqrt{AB}e^{c_1t}\cos nx}{B + 2\sqrt{AB}e^{c_1t}\cos nx + Ae^{2c_1t}}.$$

If we take $\alpha = B = 1$, $A = \epsilon^2$, and $n = N$, then we get the same solution as obtained

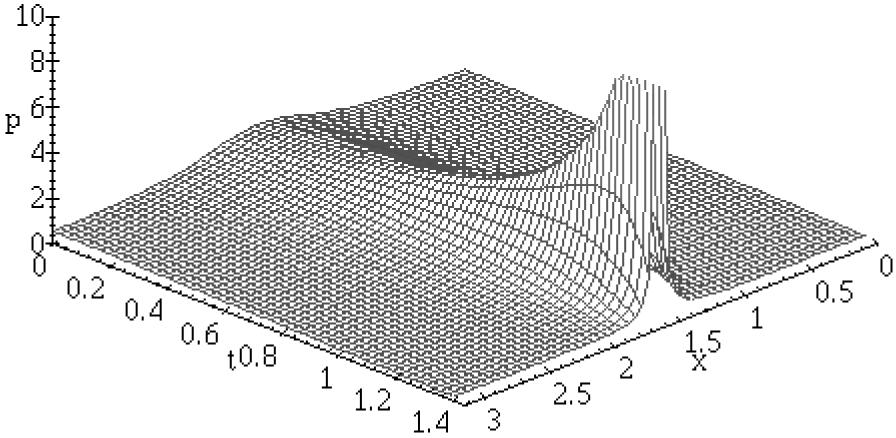


FIG. 3. *In the case of Theorem* 3.1(b)*, for the solution* $p(x,t)$*, where* $A = 0.1$*,* $B = 1$*,* $c_1 = 0.84$*,* $\alpha = 1$*,* $n = 2$*.*

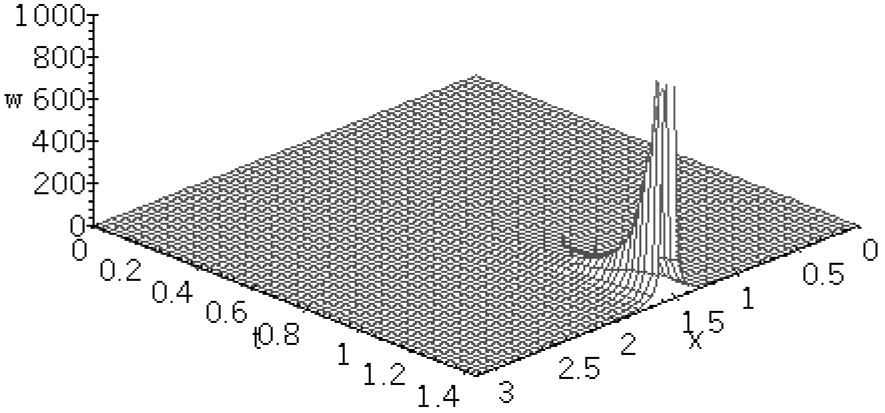


FIG. 4. *In the case of Theorem* 3.1(b)*, for the solution* $w(x,t)$*, where* $A = 0.1$*,* $B = 1$*,* $c_1 = 0.84$*,* $\alpha = 1$*,* $n = 2$*.*

by Levine and Sleeman [2].

Remark 3.6. For the problem (3.2), we can consider $\psi(x,t) = -\ln(c - \phi(x,t))$ directly, and so we get

$$\begin{cases} p(x,t) = k_1 - \mu \dfrac{Ae^{\mu t} - Be^{-\mu t}}{Ae^{\mu t} + Be^{-\mu t} \pm 2\sqrt{AB}\cos nx}, \\ w(x,t) = \dfrac{e^{k_1 t}}{Ae^{\mu t} + Be^{-\mu t} \pm 2\sqrt{AB}\cos nx}, \end{cases}$$

where $\mu = n\sqrt{k_1}$. It is clear that the solution $(p(x,t), w(x,t))$ blows up if $B > A > 0$. Furthermore, the blow-up time is $T = \frac{1}{\mu}\ln(\sqrt{\frac{B}{A}})$, and the blow-up point is $x_0 \in [0,\pi]$ satisfying $\cos nx_0 = \mp 1$. $(p(x,t), w(x,t))$ is global if $A > B > 0$ and $k_1 > n^2\{\frac{\sqrt{A}+\sqrt{B}}{\sqrt{A}-\sqrt{B}}\}^2$.

By the arguments above, we know that the maximal existence time of the solution for the system (3.1) is strongly dependent on the choice for their initial data. Also, we can investigate the asymptotic behavior of the solution and find something very interesting: First, we can get two pairs of solutions $(p_i(x,t), w_i(x,t))(i = 1, 2)$ such that even though the difference of their initial data is relatively large, they still have very similar asymptotic behavior:

$$\lim_{t\to+\infty}(p_1(x,t) - p_2(x,t)) = 0, \qquad \lim_{t\to+\infty}(w_1(x,t) - w_2(x,t)) = 0 \qquad \text{for } x \in \bar{\Omega}.$$

Second, even if the difference of their initial data is very small, we can also find two solutions $(p_i(x,t), w_i(x,t))$, $i = 1, 2$, in which one solution exists globally and another one will blow up in finite time. This means that these biological systems are unstable.

We have following result.

THEOREM 3.2. *For any positive spatial independent solution of the dynamics* (3.1) $(p_0, w_0e^{p_0t})$, *we have the following results.*

(a) *There exists a family of positive solutions for the dynamics* (3.1)

$$\begin{cases} p(x,t) &= k - \mu\dfrac{Ae^{\mu t} - Be^{-\mu t}}{Ae^{\mu t} + Be^{-\mu t} \pm 2\sqrt{AB}\cos Nx}, \\ w(x,t) &= \dfrac{e^{kt}}{Ae^{\mu t} + Be^{-\mu t} \pm 2\sqrt{AB}\cos Nx}, \end{cases}$$

where $\mu = N\sqrt{k}$, *satisfying*

$$\lim_{t\to+\infty}(p(x,t) - p_0) = 0, \quad \lim_{t\to+\infty}\left(\frac{w_0e^{p_0t}}{w(x,t)}\right) = 1.$$

(b) *For any positive constant* $\varepsilon > 0$ *small enough, we can find a positive solution* $(p(x,t), w(x,t))$ *with the initial data* $p_0 - \varepsilon < p(x,0) < p_0 + \varepsilon$ *and* $0 < w(x,0) < \varepsilon$, *and* $(p(x,t), w(x,t))$ *will blow up at finite-time* T.

Proof. (a) For any $p_0 > 0, w_0 > 0$, and positive integer $N > 0$ fixed, denote by $\sqrt{k}$ the positive solution of the equation

$$\lambda^2 - N\lambda - p_0 = 0.$$

Then we have $k - N\sqrt{k} = p_0 > 0$. Since the function $\frac{\sqrt{A}+x}{\sqrt{A}-x}$ is strictly increasing continuous for $0 < x < \sqrt{A}$, satisfying

$$\lim_{x\to0^+}\frac{\sqrt{A}+x}{\sqrt{A}-x} = 1, \quad \lim_{x\to\sqrt{A}^-}\frac{\sqrt{A}+x}{\sqrt{A}-x} = +\infty$$

for any positive constant A, we can find $B > 0$ sufficiently small such that

$$k > N^2\left\{\frac{\sqrt{A}+\sqrt{B}}{\sqrt{A}-\sqrt{B}}\right\}^2.$$

According to Theorem 3.1, the solution of the dynamics (3.1)

$$\begin{aligned} p(x,t) &= k - N\sqrt{k}\frac{Ae^{N\sqrt{k}t} - Be^{-N\sqrt{k}t}}{Ae^{N\sqrt{k}t} + Be^{-N\sqrt{k}t} \pm 2\sqrt{AB}\cos Nx}, \\ w(x,t) &= \frac{e^{kt}}{Ae^{N\sqrt{k}t} + Be^{-N\sqrt{k}t} \pm 2\sqrt{AB}\cos Nx} \end{aligned}$$

exists globally, and

$$\lim_{t\to+\infty} p(x,t) = k - N\sqrt{k} = p_0, \quad \lim_{t\to+\infty}\left[\frac{w_0 e^{p_0 t}}{w(x,t)}\right] = Aw_0.$$

If we choose $A = \frac{1}{w_0}$, we can get the result as mentioned in (a).

(b) For any $\varepsilon > 0$, we can find a positive constant $B_0 > 0$ sufficiently large such that for any $B \geq B_0$, $|p_0(\frac{\sqrt{B}-\sqrt{A}}{\sqrt{B}+\sqrt{A}} - 1)| < \varepsilon$, $|p_0(\frac{\sqrt{B}+\sqrt{A}}{\sqrt{B}-\sqrt{A}} - 1)| < \varepsilon$ for some positive constant A, where we suppose $B_0 > A$. Let $k = \frac{(\sqrt{N^2+4p_0}-N)^2}{4}$; then we have

$$k + N\sqrt{k} = p_0.$$

Choose $T > 0$ satisfying

$$T = \frac{1}{N[\sqrt{N^2+4p_0} - N]}\ln\frac{B}{A} > 0.$$

It is well known that the solution $(p(x,t), w(x,t))$, denoting $N\sqrt{k}$ by μ, where

$$\begin{aligned} p(x,t) &= k + \mu\frac{Be^{-\mu t} - Ae^{\mu t}}{Be^{-\mu t} + Ae^{\mu t} \pm 2\sqrt{AB}\cos Nx}, \\ w(x,t) &= \frac{e^{kt}}{Be^{-\mu t} + Ae^{\mu t} \pm 2\sqrt{AB}\cos Nx}, \end{aligned}$$

is a solution for the system (3.1) with the initial data

$$p_0 - \varepsilon < k + \mu\frac{\sqrt{B}-\sqrt{A}}{\sqrt{B}+\sqrt{A}} \leq p(x,0) \leq k + \mu\frac{\sqrt{B}+\sqrt{A}}{\sqrt{B}-\sqrt{A}} < p_0 + \varepsilon$$

that will blow up in finite-time T at the point x_0 chosen above.

The proof is completed. □

Remark 3.7. According to the statement above, we know that the maximum existence of the solution for the dynamics (3.1) is very sensitive to the initial data. Even close to the constant initial data, we can find both global and blow-up in finite-time solutions, respectively.

Remark 3.8. We can find from the argument in the proof of (b) of Theorem 3.2 that the blow-up time is

$$T = \frac{1}{N[\sqrt{N^2+4p_0} - N]}\ln\frac{B}{A} = \frac{[\sqrt{N^2+4p_0} + N]}{4p_0 N}\ln\frac{B}{A} = \frac{\sqrt{1+\frac{4p_0}{N^2}} + 1}{4p_0}\ln\frac{B}{A}.$$

It is obvious that T is increasing with B.

4. High dimensional case with *w* possessing exponential growth. In section 3, we knew that even in the special case, when w has exponential growth, the behavior of the solutions is very complicated. Let us consider the following general case:

$$
(4.1)\qquad \begin{cases} \dfrac{\partial p}{\partial t} = D\nabla\left(p\nabla\left(\ln\left(\dfrac{p}{w}\right)\right)\right) & \text{for } x \in \Omega \subseteq R^n, \quad t > 0,\\ \dfrac{\partial w}{\partial t} = (\beta p - \mu)w & \\ p\nabla\left(\ln\left(\dfrac{p}{w}\right)\right)\cdot \vec{n} = 0 & \text{for } x \in \partial\Omega, \quad t > 0,\\ p(x,0) = p_0(x) > 0 & \text{for } x \in \overline{\Omega},\\ w(x,0) = w_0(x) > 0 & \end{cases}
$$

where $\Omega, \vec{n}, D, \beta$, and μ are all the same as in the preceding section and $n \geq 1$. Let $u(x,t) = \frac{p(x,t)}{e^{\mu t}w(x,t)}$. Since $(e^{\mu t}w)_t = \beta e^{\mu t}wp$, $u = \frac{p}{e^{\mu t}w} = \frac{(e^{\mu t}w)_t}{\beta(e^{\mu t}w)^2}$, we have

$$
(4.2)\quad w(x,t) = \frac{w_0(x)e^{-\mu t}}{1 - \beta w_0(x)\displaystyle\int_0^t u(x,\tau)\,d\tau}, \qquad p(x,t) = \frac{w_0(x)u(x,t)}{1 - \beta w_0(x)\displaystyle\int_0^t u(x,\tau)\,d\tau}.
$$

Also, we easily find that $(p(x,t), w(x,t))$ is a solution of the dynamics (4.1) if and only if $(u(x,t), w(x,t))$ is a solution of the following initial boundary-value problem:

$$
\begin{cases} \dfrac{\partial u}{\partial t} = D\Delta u + D\dfrac{1}{w}(\nabla w)\cdot(\nabla u) - \dfrac{\beta w_0 u^2}{1 - \beta w_0\displaystyle\int_0^t u(x,\tau)\,d\tau} & \text{for } x \in \Omega, \quad t > 0,\\ w(x,t) = \dfrac{w_0(x)e^{-\mu t}}{1 - \beta w_0(x)\displaystyle\int_0^t u(x,\tau)\,d\tau} & \\ \dfrac{\partial u}{\partial n} = 0 & \text{for } x \in \partial\Omega, \quad t > 0,\\ u(x,0) = u_0(x) = \dfrac{p_0(x)}{w_0(x)} > 0 & \text{for } x \in \overline{\Omega}.\\ w(x,0) = w_0(x) > 0 & \end{cases}
$$
(4.3)

Similarly as in the preceding section, we know that there is a unique smooth positive solution for smooth initial data locally in time. What we want to discuss is the solutions of the systems (4.1) that do not exist globally. It is worthwhile to point out that the space-independent solution $(p_0, w_0e^{(\beta p_0 - \mu)_t})$, p_0, w_0 both positive constants, is a global solution of the dynamics (4.1). In the following, we consider only the solution (p, w) that is not independent of the spatial variable. In fact, let

$$
T = \sup\{\tilde{T} > 0;\ (p(x,t), w(x,t)) \text{ exists in } \Omega \times (0, \tilde{T}]\};
$$

then by using the maximum principle, for the solution $(u(x,t), w(x,t))$ of the dynamics (4.3), we can easily obtain

$$
(4.4)\qquad 0 < u(x,t) < \max_{x\in\overline{\Omega}} u_0(x), \quad w(x,t) > 0.
$$

Also observe that, for $x \in \Omega$, $0 < t < T$, we have

$$
(4.5)\qquad 0 < \beta w_0(x)\int_0^t u(x,\tau)d\tau \leq \max_{x\in\overline{\Omega}}\left[\beta w_0(x)\int_0^t u(x,\tau)d\tau\right] < 1.
$$

Let us define

$$\underline{w}_0 \overset{\triangle}{=} \min_{x\in\overline{\Omega}} w_0(x), \quad \overline{w}_0 \overset{\triangle}{=} \max_{x\in\overline{\Omega}} w_0(x),$$

$$\underline{u}_0 \overset{\triangle}{=} \min_{x\in\overline{\Omega}} u_0(x), \quad \overline{u}_0 \overset{\triangle}{=} \max_{x\in\overline{\Omega}} u_0(x),$$

$$m(t) \overset{\triangle}{=} 1 - \max_{x\in\overline{\Omega}} \left[\beta w_0(x) \int_0^t u(x,\tau)d\tau\right]$$

and

$$0 < \phi(t) \overset{\triangle}{=} \min_{x\in\overline{\Omega}} u(x,t) \le \max_{x\in\o verline\Omega} u(x,t) \overset{\triangle}{=} \psi(t) \text{ for } 0 < t < T.$$

Thus if $(p(x,t), w(x,t))$ is a global solution of (4.1), then we have the following necessary condition.

THEOREM 4.1. *If $T = +\infty$, then $m(+\infty) = 0$, which implies that*

$$\max_{x\in\overline{\Omega}} \left[\beta w_0(x) \int_0^\infty u(x,\tau)d\tau\right] = 1.$$

Proof. Since $T = +\infty$, the dynamics (4.1) has a unique global positive solution $(p(x,t), w(x,t))$ with positive initial data $p(x,0) = p_0(x)$, $w(x,0) = w_0(x)$ on $\overline{\Omega}$, which implies that the function transformation, given by (4.2), is well defined for all time and that $1/w$ is positive and finite for all time. Because $u(x,t)$ is a solution of the dynamics (4.3), we have

$$\frac{\partial u}{\partial t} - D \triangle u - D\frac{1}{w}(\nabla w)\cdot(\nabla u) = -\frac{\beta w_0(x)u^2}{1-\beta w_0(x)\displaystyle\int_0^t u(x,\tau)d\tau} \ge -\frac{\beta\overline{w}_0 u^2}{m(t)}$$

for $x \in \Omega,\ t > 0$. This implies that $u(x,t)$ is a supersolution of the following dynamics:

$$(4.6) \quad \begin{cases} \dfrac{\partial v}{\partial t} - D\triangle v - D\dfrac{1}{w}(\nabla w)\cdot(\nabla v) = -\dfrac{\beta\overline{w}_0 v^2}{m(t)} & \text{for } x\in\Omega, \quad t>0, \\ \dfrac{\partial v}{\partial \mathbf{n}} = 0 & \text{for } x \in \partial\Omega, \quad t > 0, \\ v(x,0) = \underline{u}_0 & \text{for } x \in \overline{\Omega}. \end{cases}$$

We can solve the dynamics (4.6), which has a unique solution as follows:

$$v(x,t) = v(t) = \frac{\underline{u}_0}{1+\displaystyle\int_0^t \frac{\beta\overline{w}_0\underline{u}_0}{m(t_1)}dt_1}, \quad x\in\Omega, \quad t>0;$$

thus we obtain

$$(4.7) \quad u(x,t) \ge \phi(t) \ge v(t) = \frac{\underline{u}_0}{1+\displaystyle\int_0^t \frac{\beta\overline{w}_0\underline{u}_0}{m(t_1)}dt_1}, \quad x\in\Omega, \quad t>0.$$

Observe that for all $t > 0$ we have

$$\begin{aligned}
1 &> \max_{x\in\overline{\Omega}}\left[\beta w_0\int_0^t u(x,\tau)d\tau\right] \geq \beta\overline{w}_0\int_0^t \phi(\tau)d\tau \geq \beta\overline{w}_0\int_0^t v(\tau)d\tau\\
&= \int_0^t \frac{\beta\overline{w}_0\underline{u}_0}{1+\displaystyle\int_0^{t_1}\frac{\beta\overline{w}_0\underline{u}_0}{m(t_2)}dt_2}dt_1\\
&= \int_0^t \frac{\beta\overline{w}_0\underline{u}_0}{m(t_1)}\cdot\frac{m(t_1)}{1+\displaystyle\int_0^{t_1}\frac{\beta\overline{w}_0\underline{u}_0}{m(t_2)}dt_2}dt_1\\
&\geq m(t)\int_0^t \frac{\beta\overline{w}_0(x)\underline{u}_0}{m(t_1)}\cdot\frac{1}{1+\displaystyle\int_0^{t_1}\frac{\beta\overline{w}_0\underline{u}_0}{m(t_2)}dt_2}dt_1\\
&= m(t)\ln\left(1+\int_0^t\frac{\beta\overline{w}_0\underline{u}_0}{m(t_1)}dt_1\right).
\end{aligned}$$

Keeping in mind that $\max_{x\in\overline{\Omega}}[\beta w_0(x)\int_0^t u(x,\tau)d\tau] < 1$ and $u(x,t) > 0$ for $x \in \Omega$ and $t > 0$, we can obtain, from the estimate above, that

$$m(+\infty) = 0, \text{ which implies that } \max_{x\in\overline{\Omega}}\left[\beta w_0(x)\int_0^{+\infty}u(x,\tau)d\tau\right] = 1,$$

as required. □

Theorem 4.2. *For the dynamics* (4.1), *if* (p, w) *does not exist globally, which means that* $T < +\infty$, *then the solution* $(p(x,t), w(x,t))$ *will blow up as* t *tends to* T. *Furthermore, we can find an* $x^* \in \overline{\Omega}$ *such that*

$$\lim_{t\to T^-} w(x^*,t) = +\infty \textit{ and } \lim_{t\to T^-} p(x^*,t) = +\infty.$$

Proof. Let $u(x,t) = \frac{p(x,t)}{w(x,t)e^{\mu t}}$. Then $(u(x,t), w(x,t))$ is the solution of the following system:

$$\begin{cases}
\dfrac{\partial u}{\partial t} = D\Delta u + D\dfrac{1}{w}(\nabla w)\cdot(\nabla u) - \dfrac{\beta w_0 u^2}{1-\beta w_0\displaystyle\int_0^t u(x,\tau)\,d\tau} & \\
 & \text{for } x\in\Omega, \quad 0<t<T,\\
w(x,t) = \dfrac{w_0(x)e^{-\mu t}}{1-\beta w_0(x)\displaystyle\int_0^t u(x,\tau)\,d\tau} & \\
\dfrac{\partial u}{\partial n} = 0 & \text{for } x\in\partial\Omega, \quad 0<t<T,\\
u(x,0) = u_0(x) = \dfrac{p_0(x)}{w_0(x)} > 0 & \\
 & \text{for } x\in\overline{\Omega}.\\
w(x,0) = w_0(x) > 0 &
\end{cases}$$

Let $f(x,t,u) = \frac{\beta w_0(x)}{1-\beta w_0(x)\int_0^t u(x,\tau)\,d\tau}$, and let $f_k(x,t,u) = \min\{f, k\}$. For w fixed, we denote $u_{(k)}(x,t)$ the solution of the following dynamics:

$$\text{(4.8)}\quad \begin{cases} \dfrac{\partial u}{\partial t} - D \bigtriangleup u - D\dfrac{1}{w}(\nabla w)\cdot(\nabla u) + f_k(x,t,u)u^2 = 0 & \text{for } (x,t)\in\Omega\times(0,T),\\ \dfrac{\partial u}{\partial \mathbf{n}} = 0 & \text{for } (x,t)\in\partial\Omega\times(0,T),\\ u(x,0) = u_0(x) & \text{for } x\in\overline{\Omega}. \end{cases}$$

We know that the dynamics (4.8) has a unique solution $u_{(k)}(x,t)$ in $\Omega\times(0,T)$ and $u(x,t)\le u_{(k)}(x,t)$ for all $(x,t)\in\Omega\times(0,T)$.

For any fixed k we can show that $u(x,t)\not\equiv u_{(k)}(x,t)$. Assume it is false, and assume we can find some k satisfying $u(x,t)\equiv u_{(k)}(x,t)$ for $(x,t)\in\Omega\times(0,T)$. Then $f(x,t,u)=f_k(x,t,u)\le k$. Since $w(x,t)=\frac{1}{\beta e^{\mu t}}f(x,t,u)=\frac{1}{\beta e^{\mu t}}f_k(x,t,u)$, $p(x,t)=\frac{1}{\beta}f(x,t,u)u=\frac{1}{\beta}f_k(x,t,u)u$, $w(x,T)$, $p(x,T)$ can be well defined for all $x\in\bar{\Omega}$. Replace the initial data $(p_0(x),w_0(x))$ of the dynamics (4.1) by $(p(x,T),w(x,T))$; we can extend the domain of the existence of the solution for the dynamics (4.1) from $\Omega\times(0,T)$ to $\Omega\times(0,T+\delta)$ for some $\delta>0$. That is in contradiction with the definition of T. So $u(x,t)\not\equiv u_{(k)}(x,t)$ for all k.

Notice that $w(x,t)=\frac{1}{\beta e^{\mu t}}f(x,t,u)$ for $(x,t)\in\bar{\Omega}\times(0,T)$, and for all k there is $(x_k,t_k)\in\bar{\Omega}\times(0,T)$ such that $f(x_k,t_k,u)>k$. Then $\lim_{k\to\infty}w(x_k,t_k)=+\infty$. It is obvious that $t_k\to T^-$. We can choose $\{x_{k_j}\}$ such that $x_{k_j}\to x^*\in\bar{\Omega}$; then $\lim_{t\to T^-}w(x^*,t)=+\infty$.

Since $\beta p=\frac{(e^{\mu t}w)_t}{e^{\mu t}w}$, we have

$$\beta\int_0^t p(x^*,\tau)\,d\tau=\int_0^t\frac{(e^{\mu t}w)_\tau}{e^{\mu\tau}w}\,d\tau=\ln\left(\frac{e^{\mu t}w(x^*,t)}{w_0(x^*)}\right)\overset{t\to T^-}{\longrightarrow}+\infty.$$

Noticing $T<+\infty$, we have $\lim_{t\to T^-}p(x^*,t)=+\infty$. This completes our proof. □

COROLLARY 4.1. *If the solution $(p(x,t),w(x,t))$ of the systems* (4.1) *blows up in finite time, then both $p(x,t)$ and $w(x,t)$ will blow up in the same point in the same time.*

Proof. If the solution of the systems (4.1) will blow up in finite time, say, in the time $t=T$, then at least one of functions $p(x,t)$ and $w(x,t)$ blows up at time T. If $w(x,t)$ blow up at the point $x=x_0$, we know that $p(x,t)$ will blow up at the point $x=x_0$ at the same time from the proof of Theorem 4.2 directly. Assume that $p(x,t)$ blows up. From the fact that $u(x,t)=\frac{p(x,t)}{w(x,t)e^{\mu t}}$ and

$$\begin{aligned} 0= {} & \frac{\partial u}{\partial t}-D\bigtriangleup u-D\frac{1}{w}(\nabla w)\cdot(\nabla u)-\mu u+\frac{\beta w_0(x)u^2}{1-\beta w_0(x)\displaystyle\int_0^t u(x,\tau)d\tau}\\ \ge {} & \frac{\partial u}{\partial t}-D\bigtriangleup u-D\frac{1}{w}(\nabla w)\cdot(\nabla u)-\mu u, \end{aligned}$$

and $\frac{\partial u}{\partial n}=0, u(x,0)=u_0(x)<\bar{u}_0$, we know that for the time $t=T$, there exist positive constants $M>m>0$, such that $m<u(x,t)<M$ for every $x\in\bar{\Omega}$. So we must have that $w(x_0,t)$ will blow up as time t tends to T. By the argument above, we know that the statement in the corollary is true. □

Remark 4.1. Levine and Sleeman [2] discuss the boundary condition $p_x(0,t)=p_x(\pi,t)=0$ (which is the same as the boundary conditions $w_x(0,t)=w_x(\pi,t)=0$), which is somewhat stronger than the original boundary conditions. They find that for their model (3.1), if $w_x = 0$ initially at an end point, then both w and p have to satisfy the zero-flux boundary condition on the entire existence interval.

According to our results in this paper, since $w = \frac{w_0(x)e^{-\mu t}}{1-\beta w_0(x)\int_0^t u(x,\tau)\,d\tau}$, we have

$$\begin{aligned}(\nabla w)\cdot\vec{n} &= \left\{ \frac{e^{-\mu t}}{1-\beta w_0(x)\displaystyle\int_0^t u\,d\tau}(\nabla w_0) - \frac{w_0 e^{-\mu t}}{\left(1-\beta w_0(x)\displaystyle\int_0^t u\,d\tau\right)^2} \right.\\ &\qquad \left. \times\left[-\beta(\nabla w_0)\int_0^t u\,d\tau - \beta w_0\int_0^t(\nabla u)\,d\tau\right]\right\}\cdot\vec{n}\\ &= \frac{e^{-\mu t}}{1-\beta w_0(x)\displaystyle\int_0^t u\,d\tau}\left[1+\frac{\beta w_0\displaystyle\int_0^t u\,d\tau}{1-\beta w_0\displaystyle\int_0^t u\,d\tau}\right](\nabla w_0)\cdot\vec{n}\\ &\quad + \frac{\beta w_0^2 e^{-\mu t}}{\left(1-\beta w_0\displaystyle\int_0^t u\,d\tau\right)^2}\int_0^t(\nabla u)\cdot\vec{n}\,d\tau\\ &= \frac{e^{-\mu t}}{1-\beta w_0(x)\displaystyle\int_0^t u\,d\tau}\left[1+\frac{\beta w_0\displaystyle\int_0^t u\,d\tau}{1-\beta w_0\displaystyle\int_0^t u\,d\tau}\right](\nabla w_0)\cdot\vec{n}.\end{aligned}$$

This implies that even for the original dynamics (4.1), at any point on $\partial\Omega$, $(\nabla w_0)\cdot\vec{n} = 0$ if and only if $(\nabla w(x,t))\cdot\vec{n} = 0$ on the entire existence interval. This situation also occurs for the dynamics (4.1).

5. Some results for other biological models. Othmer and Stevens have also constructed some other biological model:

$$(5.1)\qquad \begin{cases} \dfrac{\partial p}{\partial t} = D\nabla\cdot(p\nabla(\ln(pw))) & \text{for } x\in\Omega,\quad t>0,\\ \dfrac{\partial w}{\partial t} = F(p,w) & \\ p\nabla\left(\ln\left(pw\right)\right)\cdot\vec{n} = 0 & \text{for } x\in\partial\Omega,\quad t>0,\\ p(x,0) = p_0(x) > 0 & \\ w(x,0) = w_0(x) > 0 & \text{for } x\in\overline{\Omega}. \end{cases}$$

For this system we can also consider the growth rate of w as (i) $F(p,w) = \beta p - \mu w$ and (ii) $F(p,w) = (\beta p - \mu)w$; respectively. Here we introduce function transformations $u(x,t) = pw$, $u(x,t) = pwe^{\mu t}$, respectively; then we can prove the following results.

(i) When $F(p,w) = \beta p - \mu w$ and if the initial data satisfy the boundary conditions in (5.1), then there exists a unique global solution (p,w) of the dynamics (5.1) in which $p(x,t)$ is bounded. If $w_0(x) =$ constant, then there is collapse, except for the constant solution $(p,w) = (1,\frac{\beta}{\mu})c$.

(ii) When $F(p,w) = (\beta p - \mu)w$ and if the initial data satisfy the boundary conditions in (5.1), then there exists a unique global solution (p,w) of the dynamics

(5.1), and $p(x,t)$ is bounded. If $w_0(x) \leq 1$, then there is collapse, except for the constant solution $(p,w) = (\frac{\mu}{\beta}, c)$.

Since the processes of the proofs for these results are very similar to the above work in sections 2 and 4, we omit them here.

Acknowledgments. We thank B. D. Sleeman and H. A. Levine for having sent us the paper of their recent results and for many valuable and fruitful discussions. We also thank H. F. Weinberger for the comments.

The final version of the paper was completed while H. Chen was a visiting professor in the School of Mathematics, University of Leeds during September–October 1999; he also would like to thank the institute for the invitation and hospitality.

REFERENCES

[1] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasi-Linear Equations of Parabolic Type*, Amer. Math. Soc. Transl. Ser. 2, 23, AMS, Providence, RI, 1968.

[2] H. A. Levine and B. D. Sleeman, *A system of reaction diffusion equations arising in the theory of reinforced random walks*, SIAM J. Appl. Math., 57 (1997), pp. 683–730.

[3] H. G. Othmer, S. R. Dunbar, and W. Alt, *Models of dispersal in biological systems*, J. Math. Biol., 26 (1988), pp. 263–298.

[4] H. G. Othmer and A. Stevens, *Aggregation, blowup, and collapse: The ABC's of taxis and reinforced random walks*, SIAM J. Appl. Math., 57 (1997), pp. 1044–1081.

[5] M. Rascle, *On a system of non-linear strongly coupled partial differential equations arising in biology*, in Proceedings of the Dundee Conference on Ordinary and Partial Differential equations, Lecture Notes in Math. 846, W.N. Everett and B.D. Sleeman, eds., Springer-Verlag, New York, 1980, pp. 290–298.

[6] M. Rascle and C. Ziti, *Finite time blow-up in some models of chemotaxis growth*, J. Math. Biol., 33 (1995), pp. 388–414.

[7] M. Rascle, *Sur une équation intégro-différentielle non linéaire issue de la biologie*, J. Differential Equations, 32 (1979), pp. 420–453.

[8] S. Childress, *Chemotactic collapse in two dimensions,* in Modelling of Patterns in Space and Time, Lecture Notes in Biomath. 55, Springer-Verlag, Berlin, 1984, pp. 61–68.

[9] S. Childress and J. K. Percus, *Nonlinear aspects of chemotaxis*, Math. Biosci., 56 (1981), pp. 217–237.

[10] W. Jager and S. Luckhaus, *On explosions of solutions to a system of partial differential equations modelling chomtaxis*, Trans. Amer. Math. Soc., 329 (1992), pp. 819–821.

[11] M. Herrero and J. Velazquez, *Chemotactic collapse for the Keller-Segel model*, J. Math. Biol., 35 (1996), pp. 583–623.

[12] T. Nagai, *Blow-up of radially symmetric solutions to a chemotaxis system*, Adv. Math. Sci. Appl., 5 (1995), pp. 581–601.

[13] E. F. Keller and L. A. Segel, *Initiation of slime mold aggregation viewed as an instability*, J. Theor. Biol., 26 (1970), pp. 399–415.

[14] P. Biler, *Global solutions to some parabolic-elliptic systems of chemotaxis*, Adv. Math. Sci. Appl., 9 (1999), pp. 347–357.

[15] T. Nagai and T. Senba, *Global existence and blow-up of radial solutions to a parabolic-elliptic system of chemotaxis*, Adv. Math. Sci. Appl., 8 (1998), pp. 145–156.

SIAM J. MATH. ANAL.
Vol. 33, No. 4, pp. 786–789

POTATO CHIP SINGULARITIES OF 3D FLOWS*

DIEGO CORDOBA† AND CHARLES FEFFERMAN†

Abstract. A “potato chip singularity” forms when two distinct surfaces moving with a three-dimensional (3D) fluid coincide at a finite time. Potato chip singularities were suggested by a numerical study of 3D ideal magnetohydrodynamics. We prove that an incompressible flow satisfying a mild assumption on velocity growth cannot form a potato chip singularity.

Key words. incompressible flow, singularity, potato chip, magnetohydrodynamics

AMS subject classification. 76B03

PII. S0036141001384995

In this note, we show that the technique we developed in [2, 3, 4] applies to a problem posed by Grauer and Marliani [5] on the possible formation of current sheet singularities in three-dimensional (3D) ideal incompressible magnetohydrodynamics (3D MHD). Numerical simulations in [5] show rapid initial increase of the current density, leading to a large current density in a thin neighborhood of a curved surface, called a “potato chip” in [5]. If the thickness of the potato chip becomes zero in finite time, then one has a breakdown for an initially smooth 3D MHD solution. That is, the initially smooth solution cannot be continued to a smooth 3D MHD solution beyond some finite “breakdown time” T. The numerics in [5] indicate that the initial rapid growth of current density changes to a merely exponential growth. Grauer and Marliani suggest in [5] that the technique of Cordoba [1] might be used to rule out a potato chip singularity for 3D MHD. This contrasts with earlier work of Kerr and Brandenburg [6], who reported observing a breakdown for a 3D MHD solution with an initial condition similar to that of [5].

The purpose of this note is to rule out finite-time potato chip singularities for general 3D incompressible flows, under a mild assumption on the velocity growth. We begin by giving a precise definition of a potato chip singularity. Our definition applies in particular to the scenario contemplated by Grauer and Marliani [5].

Let $U = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_1^2 + x_2^2 < a^2, -\frac{H}{2} < x_3 < \frac{H}{2}\}$ be a cylinder, and let $[0, T)$ be an interval $(a, H, T > 0)$. We denote the closure of U by $\bar{U}$. A *moving surface* in $U \times [0, T)$ is defined as $S(t) = \{(x_1, x_2, x_3) \in U : x_3 = f(x_1, x_2, t)\}$ for $t \in [0, T)$, where f is a C^1 function from $\{(x_1, x_2, t) : x_1^2 + x_2^2 < a^2, 0 \le t < T\}$ to $\left(-\frac{H}{2}, \frac{H}{2}\right)$. Suppose we are given a velocity field $u(x, t) = (u_1(x, t), u_2(x, t), u_3(x, t))$, defined for $x \in \bar{U}$, $t \in [0, T)$. The moving surface $S(t)$ is said to *move with the velocity field* u if the equation

$$\frac{\partial f}{\partial t}(x_1, x_2, t) = u_1(x_1, x_2, x_3, t)\frac{\partial f}{\partial x_1}(x_1, x_2, t) + u_2(x_1, x_2, x_3, t)\frac{\partial f}{\partial x_2}(x_1, x_2, t) + u_3(x_1, x_2, x_3, t)$$

holds for all $x \in S(t)$, $t \in [0, T)$.

*Received by the editors February 11, 2001; accepted for publication (in revised form) September 4, 2001; published electronically December 7, 2001.
http://www.siam.org/journals/sima/33-4/38499.html

†Department of Mathematics, Princeton University, Fine Hall, Washington Road, Princeton, NJ 08544-1000 (dcg@math.princeton.edu, cf@math.princeton.edu). The second author’s work was partially supported by NSF grant DMS 0070692 and by the American Institute of Mathematics.

The velocity field u is said to form a *potato chip singularity* at time T if there exists a pair of moving surfaces

$$S_\pm(t) = \{(x_1, x_2, x_3) \in U : x_3 = f_\pm(x_1, x_2, t)\} \text{ in } U \times [0, T),$$

both moving with the velocity field u and satisfying the following two conditions:

- $f_-(x_1, x_2 t) < f_+(x_1, x_2, t)$ for $x_1^2 + x_2^2 < a^2$, $t \in [0, T)$,
- $\lim_{t \to T-} [f_+(x_1, x_2, t) - f_-(x_1, x_2, t)] = 0$ for $x_1^2 + x_2^2 < a^2$.

Our result on potato chip singularities is as follows.

THEOREM. *Let $u(x, t)$ be a C^1, divergence-free velocity field, defined on $\bar{U} \times [0, T)$. If we have $\int_0^T \sup_{x \in \bar{U}} |u(x, t)| dt < \infty$, then u cannot form a potato chip singularity at time T.*

This places an added burden on anyone alleging potato chip singularity formation in a numerical simulation: The velocity must be seen to grow so rapidly as to suggest the divergence of $\int_0^T \sup_{x \in U} |u(x, t)| \, dt$.

The rest of this note gives the proof of this theorem. We assume that a potato chip singularity forms at time T, and we derive a contradiction. By analogy with [3, 4], we introduce a time-varying region

$$\Omega(t) = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_1^2 + x_2^2 < (R(t))^2, \ f_-(x_1, x_2, t) < x_3 < f_+(x_1, x_2, t)\}$$

for $t \in [t_0, T)$.

Here, $0 < R(t) < a$ is an increasing C^1 function on the interval $[t_0, T)$. Both the function $R(t)$ and the initial time $t_0 \in (0, T)$ will be specified later. We will derive an obvious formula for the time derivative of the volume of $\Omega(t)$. To do so, we first note that the boundary $\partial\Omega(t)$ consists of the top and bottom,

$$E_\pm(t) = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_1^2 + x_2^2 \leq (R(t))^2, x_3 = f_\pm(x_1, x_2, t)\},$$

and the side

$$\mathcal{S}(t) = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_1^2 + x_2^2 = (R(t))^2, \ f_-(x_1, x_2, t) \leq x_3 \leq f_+(x_1, x_2, t)\}.$$

The outward-pointing unit normal ν to $\partial\Omega(t)$ is given on $E_\pm(t)$ by

$$\nu = \nu_\pm = \pm \left(1, \frac{\partial f_\pm}{\partial x_1}, \frac{\partial f_\pm}{\partial x_2}\right) \Bigg/ \sqrt{1 + \left(\frac{\partial f_\pm}{\partial x_1}\right)^2 + \left(\frac{\partial f_\pm}{\partial x_2}\right)^2}$$

and on $\mathcal{S}(t)$ by

$$\nu = \nu_0 = (x_1, x_2, 0) \Big/ \sqrt{x_1^2 + x_2^2}.$$

The derivative of the volume of $\Omega(t)$ with respect to t is given by

$$\frac{d}{dt} \text{Vol}\, \Omega(t) = \int_{x_1^2 + x_2^2 < (R(t))^2} \left(\frac{\partial f_+}{\partial t} - \frac{\partial f_-}{\partial t}\right) dx_1 dx_2 + R'(t) \int_{x_1^2 + x_2^2 = (R(t))^2} (f_+ - f_-) \, d\,(\text{length}).$$

Now we bring in the fact that the surfaces $S_\pm(t)$ move with the velocity field. We have

$$\pm\int_{x_1^2+x_2^2<(R(t))^2}\frac{\partial f_\pm}{\partial t}dx_1dx_2 = \pm\int_{E_\pm(t)}\left(u_1\frac{\partial f_\pm}{\partial x_1}+u_2\frac{\partial f_\pm}{\partial x_2}+u_3\right)dx_1dx_2$$
$$= \int_{E_\pm(t)}(u_1,u_2,u_3)\cdot\left[\pm\left(1,\,\frac{\partial f_\pm}{\partial x_1},\frac{\partial f_\pm}{\partial x_2}\right)\Big/\sqrt{1+\left(\frac{\partial f_\pm}{\partial x_1}\right)^2+\left(\frac{\partial f_\pm}{\partial x_2}\right)^2}\,\right]$$
$$\cdot\sqrt{1+\left(\tfrac{\partial f_\pm}{\partial x_1}\right)^2+\left(\tfrac{\partial f_\pm}{\partial x_2}\right)^2}\,dx_1dx_2$$
$$= \int_{E_\pm} u\cdot\nu\, d\,(\text{Area}).$$

Therefore, our previous formula for the time derivative of Vol $\Omega(t)$ may be rewritten in the form

$$\frac{d}{dt}\operatorname{Vol}\Omega(t) = \int_{E_+(t)\cup E_-(t)} u\cdot\nu d\,(\text{Area})\; +\; R'(t)\int_{\mathcal{S}(t)} d\,(\text{Area}).$$

We can rewrite this again because the velocity field u is divergence-free. In fact, the divergence theorem for $\Omega(t)$ yields

$$0 = \int_{\Omega(t)}(\nabla\cdot u)\,d\,(\text{Vol})\; = \int_{E_\pm(t)\cup E_-(t)} u\cdot\nu d\,(\text{Area})\; + \int_{\mathcal{S}(t)} u\cdot\nu\, d\,(\text{Area}).$$

Consequently, we have the following proposition.

Proposition.

$$\frac{d}{dt}\operatorname{Vol}\Omega(t)\; = \int_{\mathcal{S}(t)}[R'(t)-u\cdot\nu]\,d\,(\text{Area}).$$

This is our basic formula for the time derivative of Vol $\Omega(t)$.

Now we pick the function $R(t)$ and the starting time $t_0\in(0,T)$. We take

$$R(t)=\frac{1}{2}\,a\; - \int_t^T \sup_{y\in\bar{U}}|u(y,\tau)|\,d\tau \quad\text{for}\quad t_0\le t<T.$$

Our assumptions on u show that $R(t)$ is a C^1 function on $[t_0,T)$, with $R'(t) = \sup_{y\in\bar{U}}|u(y,t)|$, and also that $0<R(t)<a$ for all $t\in[t_0,T)$, provided t_0 is taken close enough to T. We pick t_0 to make this happen. For this particular $R(t)$, the above proposition gives

$$\frac{d}{dt}\operatorname{Vol}\Omega(t)\; = \int_{x\in\mathcal{S}(t)}\left[\sup_{y\in\bar{U}}|u(y,t)| - u(x,t)\,\cdot\,\nu(x,t)\right]d(\text{Area})\;\ge\;0$$

for all $t\in[t_0,T)$, since $\mathcal{S}(t)\subset\bar{U}$.

Since Vol $\Omega(t_0)>0$, it follows that $\liminf_{t\to T-}$ Vol $\Omega(t)>0$, and consequently

$$\liminf_{t\to T-}\int_{x_1^2+x_2^2<a^2}(f_+(x_1,x_2,t)\,-f_-(x_1,x_2,t))\,dx_1dx_2\;>\;0. \tag{1}$$

On the other hand, if a potato chip singularity forms at time T, then

$$\lim_{t\to T-}(f_+(x_1,x_2,t)\,-f_-(x_1,x_2,t))\;=\;0\quad\text{for}\quad x_1^2+x_2^2<a^2,$$

with

$$0 < f_+(x_1, x_2, t) - f_-(x_1, x_2, t) < H.$$

Therefore, the Lebesgue dominated convergence theorem implies

$$\lim_{t \to T-} \int_{x_1^2 + x_2^2 < a^2} (f_+(x_1, x_2, t) - f_-(x_1, x_2, t))\, dx_1 dx_2 = 0. \tag{2}$$

The contradiction between (1) and (2) shows that a potato chip singularity cannot form at time T. The proof of the theorem is complete.

REFERENCES

[1] D. Cordoba, *Nonexistence of simple hyperbolic blow-up for the quasi-geostrophic equation*, Ann. of Math. (2), 148 (1998), pp. 1135–1152.

[2] D. Cordoba and C. Fefferman, *Behavior of several two-dimensional fluid equations in singular scenarios*, Proc. Natl. Acad. Sci. USA, 98 (2001), pp. 4311–4312.

[3] D. Cordoba and C. Fefferman, *Scalars convected by a 2D incompressible flow*, Comm. Pure Appl. Math., to appear.

[4] D. Cordoba and C. Fefferman, *On the collapse of tubes carried by 3D incompressible flows*, Comm. Math. Phys., 222 (2001), pp. 293–298.

[5] R. Grauer and C. Marliani, *Current sheet formation in 3D ideal incompressible magnetohydrodynamics*, Phys. Rev. Lett., 84 (2000), pp. 4850–4853.

[6] R. Kerr and A. Brandenburg, *Evidence for a singularity in ideal magnetohydrodynamics: Implications for fast reconnection*, Phys. Rev. Lett., 83 (1999), pp. 1155–1158.

SIAM J. Math. Anal.
Vol. 33, No. 4, pp. 790–826

ON THE DIFFUSIVE PROFILES FOR THE SYSTEM OF COMPRESSIBLE ADIABATIC FLOW THROUGH POROUS MEDIA*

PIERANGELO MARCATI[†] AND RONGHUA PAN[‡]

Abstract. We study the Cauchy problem for the system of one dimensional compressible adiabatic flow through porous media and the related diffusive problem. We introduce a new approach which combines the usual energy methods with special L^1-estimates and use the weighted Sobolev norms to prove the global existence and large time behavior for the solutions of the problems. The asymptotic states for the solutions are given by either stationary solutions or similarity solutions depending on the behavior of the initial data when $|x| \to \infty$. Our estimates provide asymptotic time decay rates.

Key words. damping mechanism, diffusive profile, L^1-estimates, weighted energy estimates, decay rates

AMS subject classifications. 35L45, 76S05, 35Q35, 35K55

PII. S0036141099364401

1. Introduction.

The motion of the adiabatic gas flow through porous media can be modeled by the following damped hyperbolic system:

$$\begin{cases} v_t - u_x = 0, \\ u_t + p(v,s)_x = -\alpha u, \ \alpha > 0, \\ (e(v,s) + \frac{1}{2}u^2)_t + (pu)_x = -\alpha u^2. \end{cases} \tag{1.1}$$

Where v denotes the specific volume, u is the velocity, s stands for entropy, p denotes the gas pressure with $p_v(v,s) < 0$ for $v > 0$, and e is the specific internal energy for which one has $e_s \neq 0$ and $e_v + p = 0$ (due to the second law of thermodynamics). For smooth solutions, the system (1.1) is equivalent to the following one:

$$\begin{cases} v_t - u_x = 0, \\ u_t + p(v,s)_x = -\alpha u, \ \alpha > 0, \\ s_t = 0. \end{cases} \tag{1.2}$$

It is strictly hyperbolic with characteristic speeds $-\lambda_1 = \lambda_3 = \sqrt{-p_v}$ and $\lambda_2 = 0$.

In this paper, we are interested in the influence of the damping mechanism to the smoothness and the large time behavior of the solutions. We study the Cauchy problem for the system (1.2) with the following initial data:

$$(v,u,s)(x,0) = (v_0(x), u_0(x), s_0(x)), \quad x \in R, \tag{1.3}$$

satisfying the limit conditions

$$(v_0, u_0, s_0)(x) \to (v_\pm, u_\pm, s_\pm) \text{ as } x \to \pm\infty,$$

*Received by the editors December 3, 1999; accepted for publication (in revised form) October 4, 2000; published electronically December 18, 2001.
http://www.siam.org/journals/sima/33-4/36440.html

[†] Dipartmento di Matematica Pura ed Applicata, Universita degli Studi di L'Aquila, 67100, L'Aquila, Italy (marcati@univaq.it). The work of this author was supported in part by the European TMR project under contract ERBFMRX-CT 96-0033.

[‡] SISSA, Via Beirut n. 2–4, 34014 Trieste, Italy. Present address, Department of Mathematics, University of Michigan, Ann Arbor, MI 48109-1109 (panrh@math.lsa.umich.edu).

with $v_\pm > 0$. For the sake of simplicity, from now on, we take $\alpha = 1$ and $p(v, s) = (\gamma - 1)v^{-\gamma}e^s$, with $\gamma > 1$, which is the case for the polytropic gas dynamics.

The global existence with small initial data of smooth solutions for the Cauchy problem (1.2)–(1.3) has been studied first in [10] and [11] and later by [22]. Then a natural problem is the large time behavior of the solutions. From asymptotic analysis, it is known that the first term of $(1.2)_2$ decays to zero, as $t \to \infty$, faster than others. Therefore it is natural to expect that the problem (1.2)–(1.3) is time-asymptotically equivalent to the following reduced problem:

$$\begin{cases} \tilde{v}_t = -p(\tilde{v}, s)_{xx}, \\ \tilde{u} = -p(\tilde{v}, s)_x, \\ s_t = 0, \\ \tilde{v}(x, 0) = \tilde{v}_0(x), \ s(x, 0) = s_0(x), \\ \tilde{v}_0(\pm\infty) = v_\pm, \ s_0(\pm\infty) = s_\pm. \end{cases} \tag{1.4}$$

The system in (1.4) is obtained from (1.2) by approximating the momentum equation in (1.2) with Darcy's law. Since the first equation of (1.4) is parabolic, the damping mechanism in (1.2) creates some diffusive effects when t tends to infinity.

For the isentropic flow, namely, $s =$ const, (1.2) takes the following form:

$$\begin{cases} v_t - u_x = 0, \\ u_t + p(v)_x = -u. \end{cases} \tag{1.5}$$

The diffusive effect created by the damping mechanism has been investigated for the Cauchy problem of (1.5) with the initial data

$$(v(x, 0), u(x, 0)) = (v_*(x), u_*(x)) \tag{1.6}$$

such that

$$\lim_{x\to\pm\infty}(v_*(x), u_*(x)) = (v_\pm, u_\pm).$$

It has been proved in [5] that the smooth solution of (1.5)–(1.6) can be described time-asymptotically by the solution of the following parabolic problem:

$$\begin{cases} \tilde{v}_t = -p(\tilde{v})_{xx}, \\ \tilde{u} = -p(\tilde{v})_x, \\ \tilde{v}(x, 0) = \tilde{v}_*(x + d_0). \end{cases} \tag{1.7}$$

Where $\tilde{v}_*$ is the similarity solution of $(1.7)_1$ with $\tilde{v}_*(\pm\infty) = v_\pm$. For other results, we refer to [3], [4], [6], and [19] for smooth solutions and to [1], [4], [8], [12], [13], [14], [15], [17], [18], and [21] for weak solutions. For the initial boundary value problems on a quarter plane, we refer to [16] and [20].

There are few results in the literature for the case $s \neq$ const. Partial answers are given in [11] and [7] for the Cauchy problem and in the recent paper of Hsiao and Pan [9] concerning the initial boundary value problem. The case $v_- = v_+ = \bar{v}$ and $s_- = s_+ = \bar{s}$ was investigated in [11] and the case $v_- \neq v_+$ and $s_- = s_+ = \bar{s}$ was treated in [7] by using a technical condition (that they refer to in [7] as condition V) which requires us to solve the following parabolic problem

$$\begin{cases} \tilde{v}_t = -p(\tilde{v}, s)_{xx}, \\ \tilde{v}_0(x) = e^{\frac{1}{\gamma}(s(x)-\bar{s})}\tilde{v}_*(x + x_0) \end{cases} \tag{1.8}$$

and to control the behavior of its solutions by means of the similarity solutions of

$$\begin{cases} \tilde{v}_{*t} = -p(\tilde{v}_*, \bar{s})_{xx}, \\ \tilde{v}_*(\pm\infty) = v_\pm. \end{cases} \tag{1.9}$$

The purpose of this paper is to deal with the following two cases:

Case 1: $s_- = s_+ = \bar{s}$;

Case 2: $(v_\pm, s_\pm)$ satisfy $p(v_-, s_-) = p(v_+, s_+) = \bar{p}$.

In Case 1, namely, $s_- = s_+ = \bar{s}$, we cannot use the methods of [7] and we need new techniques. In particular we shall combine the usual energy methods with special L^1-estimates and with the use of weighted Sobolev norms to solve the problems in detail. This is the only case where $p(v_-, s_-) \neq p(v_+, s_+)$ that we can treat in this paper. In this case, the asymptotic states will be the similarity solution of (1.9) given by the scaling invariance with respect to the transformation $x \to \sigma x$, $t \to \sigma^2 t$. Our results strongly improved those in [7]. Indeed, we remove the technical condition V and we describe the asymptotic states both for the diffusion problem and the hyperbolic one by using the similarity solutions. Thanks to our new approach, it is possible to get a decay rate which did not exist in the previous results (see [7]). Our results on the parabolic problem generalize the result of [2] to the adiabatic case, and also obtain better decay rates.

In Case 2, we can determine a special solution $v_3(x)$ to $(1.4)_1$ by solving the equation $p(v_3, s) = \bar{p}$. Then in this case we establish results similar to those obtained in [11] with, in addition, some decay rates.

Before stating the main results, we describe the plan of this paper. In section 2, the parabolic problem (1.4) is studied in detail for both cases by using our new approach. Then sections 3 and 4 are devoted to the hyperbolic problem (1.2)–(1.3) for Case 1 and Case 2, respectively.

We now state our main results.

1.1. Main results: Parabolic equation. Since in (1.2) or (1.4) $s_t = 0$, then $s(x,t) = s(x) = s_0(x)$. Let us denote

$$\begin{aligned} a(x) &= (\gamma-1)^{-\frac{1}{\gamma}} e^{-\frac{1}{\gamma}s(x)}, \\ a_1 &= (\gamma-1)^{-\frac{1}{\gamma}} e^{-\frac{1}{\gamma}\bar{s}}, \\ w &\equiv a(x)\tilde{v} = p(\tilde{v}, s)^{-\frac{1}{\gamma}}, \end{aligned} \tag{1.10}$$

then (1.4) is equivalent to the following:

$$\begin{cases} w_t + a(x)(w^{-\gamma})_{xx} = 0, \\ \tilde{u} = -(w^{-\gamma})_x, \\ s(x,t) = s_0(x), \\ w(x,0) = w_0(x) = a(x)\tilde{v}_0(x), \\ w(\pm\infty) = w_\pm > 0. \end{cases} \tag{1.11}$$

Moreover, we will denote by $\tilde{w}(\eta)$ (with $\eta = \frac{x}{\sqrt{t+1}}$) the similarity solution of the following problem:

$$\begin{cases} \tilde{w}_t + a_1(\tilde{w}^{-\gamma})_{xx} = 0, \\ \tilde{w}(\pm\infty) = w_\pm. \end{cases} \tag{1.12}$$

By combining the weighted energy method and L^1-estimate, we can prove the following theorem for Case 1.

Theorem 1.1. *Assume that $w_0(x)$ and $s_0(x)$ are C^2 functions and*

$$w_0(x) - \tilde{w}(x,0) \in H^2(R) \cap L^1(R), \quad x(s_0(x) - \bar{s}) \in L^1(R).$$

There exists $\delta_0 > 0$ such that, if $0 < \delta < \delta_0$ and

$$|w_+ - w_-| + \|w_0(x) - \tilde{w}(x,0)\|_{H^2} \le \delta,$$

then (1.11) *has a unique global smooth solution $(w, \tilde{u}, s)(x,t)$ satisfying*

$$w(x,t) - \tilde{w} \in C^0([0,t]; H^2) \ for \ all \ t > 0.$$

Moreover, there exist positive constants $C > 0$, $\beta_1 > \frac{1}{3}$, and $\beta_2 > \frac{1}{2}$ such that

$$\|w(x,t) - \tilde{w}\|_{L^\infty} \le C(1+t)^{-\frac{1}{2}}(1 + \log(1+t))^{\beta_1},$$
$$\|\tilde{u} + (\tilde{w}^{-\gamma})_x\|_{L^\infty} \le C(1+t)^{-1}(1 + \log(1+t))^{\beta_2}.$$

Thus, by setting $\tilde{v} = a^{-1}w$, $\hat{v} = a^{-1}\tilde{w}$, and $\hat{u} = -(\tilde{w}^{-\gamma})_x$, one obtains the (unique) global smooth solution $(\tilde{v}, \tilde{u}, s)$ to (1.4) *which satisfies*

$$\|\tilde{v} - \hat{v}\|_{L^\infty} \le C(1+t)^{-\frac{1}{2}}(1 + \log(1+t))^{\beta_1},$$
$$\|\tilde{u} - \hat{u}\|_{L^\infty} \le C(1+t)^{-1}(1 + \log(1+t))^{\beta_2}.$$

Remark 1. (a) Our results in Theorem 1.1 generalize the ones in [2] to the adiabatic case and extend to a larger class of initial data. The decay rate here is better than in [2] and is almost optimal.

(b) The condition $x(s(x) - \bar{s}) \in L^1(R)$ can be replaced by the weaker one:

$$|x|^\beta (s(x) - \bar{s}) \in L^1(R)$$

for $\beta > 0$. This is clear from our proof below.

For Case 2, where $w_- = w_+ = \bar{w}$ in (1.11), it is clear that $(\bar{w}, 0, s_0(x))$ is a special solution for the system in (1.11). Let us denote $v_1(x) = a^{-1}\bar{w}$ and we have the following.

Theorem 1.2. *Assume that $w_0(x)$ and $s_0(x)$ are C^2 functions and $w_0 - \bar{w} \in H^2$. There exists $\delta_0 > 0$ such that if $0 < \delta < \delta_0$ and $\|w_0 - \bar{w}\|_{H^2} \le \delta$, then* (1.11) *has a unique global smooth solution $(w, \tilde{u}, s)(x,t)$ satisfying*

$$\lim_{t\to\infty} \|w(x,t) - \bar{w}\|_{L^\infty} = 0.$$

Furthermore, if $w_0(x) - \bar{w} \in L^1$, then

$$\|w(x,t) - \bar{w}\|_{L^\infty} \le C(1+t)^{-\frac{1}{2}}(1 + \log(1+t))^{\beta_1},$$
$$\|\tilde{u}\|_{L^\infty} \le C(1+t)^{-1}(1 + \log(1+t))^{\beta_2}.$$

Thus, by setting $v_2 = a^{-1}w$ and $u_2 = \tilde{u}$, one has a unique global smooth solution (v_2, u_2, s) to (1.4) *such that*

$$\|v_2 - v_1\|_{L^\infty} \le C(1+t)^{-\frac{1}{2}}(1 + \log(1+t))^{\beta_1},$$
$$\|u_2\|_{L^\infty} \le C(1+t)^{-1}(1 + \log(1+t))^{\beta_2}.$$

1.2. Main results: Hyperbolic problems. Based on the results in the previous theorems, we can solve (1.2)–(1.3) in detail for both cases, respectively.

Following [5], we define

$$\begin{aligned} &m(x,t) \equiv -(u_+ - u_-)m_0(x)e^{-t}, \\ &u_m(x,t) \equiv u_- e^{-t} + \int_{-\infty}^{x} m_t(\xi,t)\, d\xi, \end{aligned} \tag{1.13}$$

where $m_0(x)$ is a smooth function with compact support such that

$$\int_{-\infty}^{+\infty} m_0(x)\, dx = 1.$$

We first treat Case 1, where $s_- = s_+ = \bar{s}$. Denote by $(\tilde{v}, \tilde{u}, s)$ the solution to (1.4) obtained in Theorem 1.1. In addition, we assume

$$\int_{-\infty}^{+\infty} (v_0(x) - \tilde{v}_0(x))\, dx = -(u_+ - u_-). \tag{1.14}$$

A special choice of $\tilde{v}_0$ is given in Remark 2 below. Let us denote $y(x,t) = \int_{-\infty}^{x}(v - \tilde{v} - m)(\xi,t)\, d\xi$, then y satisfies

$$\begin{cases} y_{tt} + [p(y_x + \tilde{v} + m, s) - p(\tilde{v}, s)]_x + y_t = p(\tilde{v}, s)_{xt}, \\ y(x,0) = y_0(x) = \displaystyle\int_{-\infty}^{x} (v_0(\xi) - \tilde{v}_0(\xi) - m(\xi,0))\, d\xi, \\ y_t(x,0) = y_1(x) = u_0(x) - \tilde{u}(x,0) - u_m(x,0). \end{cases} \tag{1.15}$$

Theorem 1.3. *Under the conditions of Theorem* 1.1, *there exists $\varepsilon_0 > 0$ such that for all $0 < \varepsilon < \varepsilon_0$ and $\|y_0\|_{H^3} + \|y_1\|_{H^2} \le \varepsilon$, the system* (1.15) *admits a unique global smooth solution y such that*

$$y \in C^0([0,t]; H^3), \quad y_t \in C^0([0,t]; H^2)$$

for all $t > 0$. Moreover, there exists $C = C(\varepsilon) > 0$ such that

$$\|y_x\|_{L^\infty} \le C(1+t)^{-\frac{3}{4}}, \ \|y_t\|_{L^\infty} \le C(1+t)^{-\frac{5}{4}}.$$

Hence, by setting $v(x,t) = \tilde{v} + m + y_x$ and $u(x,t) = \tilde{u} + u_m + y_t$, one has the (unique) global smooth solution (v,u,s) to (1.2)–(1.3), *such that*

$$\|v - \tilde{v}\|_{L^\infty} \le C(1+t)^{-\frac{3}{4}}, \ \|u - \tilde{u}\|_{L^\infty} \le C(1+t)^{-\frac{5}{4}}.$$

Furthermore, in view of Theorem 1.1, *one has*

$$\begin{aligned} \|v - \hat{v}\|_{L^\infty} &\le C(1+t)^{-\frac{1}{2}}(1 + \log(1+t))^{\beta_1}, \\ \|u - \hat{u}\|_{L^\infty} &\le C(1+t)^{-1}(1 + \log(1+t))^{\beta_2}, \end{aligned}$$

where β_1 and β_2 are the same as before.

Remark 2. (a) The global existence for the smooth solution to (1.2)–(1.3) has been proved in [22], via characteristic method, provided that the initial data are small. We present here an alternative version in H^2 spaces by the energy estimate method.

(b) (1.14) is the restriction on the initial data which comes from (1.2)–(1.3) and (1.4). This is also the case for (1.16) and (1.18) below. There is a large class of functions $\tilde{v}_0(x)$ which can be chosen (for any given $v_0(x)$ in (1.3)). A special choice is $\tilde{v}_0(x) = a^{-1}\tilde{w}(x + x_0, 0)$, where x_0 is uniquely determined by

$$\int_{-\infty}^{+\infty} (v_0(x) - a^{-1}\tilde{w}(x + x_0, 0))\ dx = -(u_+ - u_-),$$

which is the special case discussed in [7]. Since Theorem 1.1 is obviously valid for $w_0(x) = \tilde{w}(x + x_0, 0)$, we include the results of [7].

(c) A similar condition was used in [9] for the initial boundary problem related to (1.2), where it was proved that both the solutions to the damped hyperbolic problem and those of the related diffusive problem have the same time asymptotic states if the initial total excessive mass is zero.

Let us now consider Case 2. Assume that

$$\int_{-\infty}^{+\infty} (v_0(x) - v_2(x, 0))\ dx = -(u_+ - u_-), \tag{1.16}$$

and denote by $\tilde{y}(x, t) = \int_{-\infty}^{x} (v - v_2 - m)(\xi, t)\ d\xi$, then we have

$$\begin{cases} \tilde{y}_{tt} + [(p(\tilde{y}_x + v_2 + m, s) - p(v_2, s)]_x = p(v_2, s)_{xt}, \\ \tilde{y}(x, 0) = \tilde{y}_0(x) = \int_{-\infty}^{x} (v_0(\xi) - v_2(\xi, 0) - m(\xi, 0))\ d\xi, \\ \tilde{y}_t(x, 0) = \tilde{y}_1(x) = u_0(x) - u_2(x, 0) - u_m(x, 0). \end{cases} \tag{1.17}$$

Similarly to Theorem 1.3, we have the following.

THEOREM 1.4. *Under the hypotheses of the Theorem* 1.2*, there exists* $\varepsilon_0 > 0$ *such that if* $0 < \varepsilon < \varepsilon_0$ *and* $\|\tilde{y}_0\|_{H^3} + \|\tilde{y}_1\|_{H^2} \le \varepsilon$*, then* (1.17) *has a unique global smooth solution* $\tilde{y}$ *such that*

$$\tilde{y} \in C^0([0, t]; H^3),\ \ \tilde{y}_t \in C^0([0, t]; H^2)$$

for all $t > 0$*. Moreover, there exists* $C = C(\varepsilon) > 0$ *such that*

$$\|\tilde{y}_x\|_{L^\infty} \le C(1 + t)^{-\frac{3}{4}},\ \ \|\tilde{y}_t\|_{L^\infty} \le C(1 + t)^{-\frac{5}{4}}.$$

Hence, by setting $v(x, t) = v_2 + m + \tilde{y}_x$ *and* $u(x, t) = u_2 + u_m + \tilde{y}_t$*, one has the (unique) global smooth solution* (v, u, s) *to* (1.2)–(1.3)*; moreover,*

$$\|v - v_2\|_{L^\infty} \le C(1 + t)^{-\frac{3}{4}},\ \ \|u - u_2\|_{L^\infty} \le C(1 + t)^{-\frac{5}{4}}.$$

Furthermore, in view of Theorem 1.2*, one has*

$$\begin{aligned} &\|v - v_1\|_{L^\infty} \le C(1 + t)^{-\frac{1}{2}}(1 + \log(1 + t))^{\beta_1}, \\ &\|u\|_{L^\infty} \le C(1 + t)^{-1}(1 + \log(1 + t))^{\beta_2}. \end{aligned}$$

Since $(v_1(x), 0, s(x))$ is a special solution to both (1.2) and (1.4), if we assume

$$\int_{-\infty}^{+\infty} (v_0 - v_1)(x)\ dx = -(u_+ - u_-) \tag{1.18}$$

and denote by $z(x,t) = \int_0^x (v - v_1 - m)(\xi, t)\ d\xi$, we have

$$(1.19) \qquad \begin{cases} z_{tt} + [p(z_x + v_1 + m, s) - p(v_1, s)]_x + z_t = 0, \\ z(x,0) = z_0(x) = \displaystyle\int_{-\infty}^{x} (v_0(\xi) - v_1(\xi) - m(\xi, 0))\ d\xi, \\ z_t(x,0) = z_1(x) = u_0(x) - u_m. \end{cases}$$

Then in this special case, we have the following result, which includes the paper [11].

Theorem 1.5. *There exists $\varepsilon_0 > 0$ such that if $0 < \varepsilon < \varepsilon_0$ and $\|z_0\|_{H^3} + \|z_1\|_{H^2} \leq \varepsilon$, then* (1.19) *has a unique global smooth solution z such that*

$$z \in C^0([0,t]; H^3),\ \ z_t \in C^0([0,t]; H^2)$$

for all $t > 0$. Moreover,

$$\|z_x\|_{L^\infty} \leq C(1+t)^{-\frac{3}{4}},\ \ \|z_t\|_{L^\infty} \leq C(1+t)^{-\frac{5}{4}}.$$

Hence, by setting $v(x,t) = v_1 + m + z_x$ and $u(x,t) = u_m + z_t$, one has the (unique) global smooth solution (v, u, s) to (1.2)–(1.3) *such that*

$$\|v - v_1\|_{L^\infty} \leq C(1+t)^{-\frac{3}{4}},\ \ \|u\|_{L^\infty} \leq C(1+t)^{-\frac{5}{4}}.$$

We will end this introduction by making a reduction. In fact, in sections 3 and 4, we will only prove Theorems 1.3–1.5 for the case $u_- = u_+ = 0$, where $m(x,t) = 0$ and $u_m = 0$. The general case can be treated in the similar way since $m(x,t)$ and u_m decay to zero exponentially fast.

2. Nonlinear diffusion equation.

This section is devoted to studing the diffusive problem (1.4). Clearly one has $s(x,t) = s_0(x) \equiv s(x)$ for all $t > 0$, which then is sufficient to solve the following equation:

$$(2.1) \qquad \begin{cases} \tilde{v}_t = -p(\tilde{v}, s)_{xx}, \\ \tilde{v}(x,0) = \tilde{v}_0(x),\ \tilde{v}_0(\pm\infty) = v_\pm > 0. \end{cases}$$

The equation (2.1) is equivalent to the following porous media type equation:

$$(2.2) \qquad \begin{cases} w_t + a(x)(w^{-\gamma})_{xx} = 0, \\ w(x,0) = w_0(x) = a(x)\tilde{v}_0(x), \\ w(\pm\infty) = w_\pm > 0, \end{cases}$$

where $a(x) = (\gamma - 1)^{-\frac{1}{\gamma}} e^{-\frac{1}{\gamma} s(x)}$, $w \equiv a(x)\tilde{v} = p(\tilde{v}, s)^{-\frac{1}{\gamma}}$. We will study the equation (2.2) instead of (2.1) for the following two cases, which are equivalent to those stated in the introduction.

Case 1: $s_- = s_+ = \bar{s}$.

Case 2: $(v_\pm, s_\pm)$ are chosen such that $w_- = w_+ = \bar{w}$, where we set $w_\pm = w(v_\pm, s_\pm)$.

We will concentrate our main efforts on Case 1 which is the most difficult part.

2.1. Case 1: $s_- = s_+ = \bar{s}$. In this subsection, (2.2) will be solved near the similarity solution for the related isentropic problem.

Now let us recall some results on the similarity solution for $(2.2)_1$ with $s(x) = \mathrm{const} = \bar{s}$. In this case, $(2.2)_1$ takes the form

$$w_t + a_1(w^{-\gamma})_{xx} = 0, \tag{2.3}$$

with $a_1 = (\gamma-1)^{-\frac{1}{\gamma}}e^{-\frac{1}{\gamma}\bar{s}}$. It is well known that (2.3) has a unique (up to a shift) similarity solution $\tilde{w}(\eta)$ (where $\eta = \frac{x}{\sqrt{1+t}}$) satisfying the limiting conditions $\tilde{w}(\pm\infty) = w_\pm$. Some properties of $\tilde{w}(\eta)$ are listed in the following lemma (see, for instance, [5]).

LEMMA 2.1. *Let $\tilde{w}(\eta)$ be the similarity solution to* (2.3) *with $\tilde{w}(\pm\infty) = w_\pm$ and $\eta = \frac{x}{\sqrt{1+t}}$. It follows that*

$$|\tilde{w}'(\eta)| + |\tilde{w}''(\eta)| \le C_1|w_+ - w_-|\exp\{-C_2\eta^2\},$$

$$|\tilde{w}(\eta) - w_-|_{\eta<0} + |\tilde{w}(\eta) - w_+|_{\eta>0} \le C_1|w_+ - w_-|\exp\{-C_2\eta^2\},$$

$$\tilde{w}_x = (1+t)^{-\frac{1}{2}}\tilde{w}'(\eta),\ \tilde{w}_t = -\frac{1}{2}(1+t)^{-1}\eta\tilde{w}'(\eta),\ (\tilde{w}^{-\gamma})_{xx} = -a_1^{-1}\tilde{w}_t,$$

$$\|D_t^iD_x^j\tilde{w}(\cdot,t)\|^2 \le C|w_+ - w_-|^2(1+t)^{-(2i+j)+\frac{1}{2}},$$

$$\|D_t^iD_x^j\tilde{w}(\cdot,t)\|_{L^\infty} \le C_1|w_+ - w_-|(1+t)^{-(i+\frac{1}{2}j)}$$

for $i+j \ge 1$ and $i \ge 0,\ j \ge 0$.

We now prove Theorem 1.1 by comparing $w(x,t)$ with $\tilde{w}(\eta)$.

Let us denote $\phi = w - \tilde{w}$; then from (2.2) and (2.3) we have the following equation:

$$\begin{cases}\phi_t + a(x)(\psi(\tilde{w})\phi)_{xx} + (a-a_1)(\tilde{w}^{-\gamma})_{xx} + a(x)(g(\phi,\tilde{w})\phi^2)_{xx} = 0,\\ \phi(x,0) = \phi_0(x) = w_0(x) - \tilde{w}(x,0).\end{cases} \tag{2.4}$$

Here

$$\psi(\tilde{w}) = -\gamma\tilde{w}^{-(\gamma+1)}$$

$$g(\phi,\tilde{w})\phi^2 = (\phi+\tilde{w})^{-\gamma} - \tilde{w}^{-\gamma} - \psi(\tilde{w})\phi.$$

Now let $F = -\psi(\tilde{w})\phi$; the corresponding problem on F is given by

$$\begin{cases}F_t + a(x)\psi(\tilde{w})F_{xx} - \psi(\tilde{w})(a-a_1)(\tilde{w}^{-\gamma})_{xx}\\ \quad -\psi_1(\tilde{w})F\tilde{w}_t - a\psi(\tilde{w})(fF^2)_{xx} = 0,\\ F(x,0) = F_0(x) = -\psi(\tilde{w}(x,0))\phi_0(x),\end{cases} \tag{2.5}$$

where

$$-\psi_1(\tilde{w})F = \psi'(\tilde{w})\phi,\ fF^2 = g\phi^2.$$

We will establish the global existence and large time behavior, for the solution F to (2.5), in the Banach space $X(0,T)$ defined for all $T > 0$ by

$$X(0,t) = \{F \in C^0([0,t];H^2),\quad 0 \le t \le T\}$$

and equipped with the norm

$$N^2(t) = \sup_{0 \le \tau \le t} \|F(\tau)\|_{H^2}^2.$$

The main result of this subsection is the following theorem.

THEOREM 2.2. *Assume that* $F_0(x)$ *and* $s(x) = s_0(x)$ *are* C^2 *functions such that* $F_0 \in H^2(R) \cap L^1(R)$ *and*

$$x(s(x) - \bar{s}) \in L^1(R). \tag{2.6}$$

Then there exist constants $\varepsilon_0 > 0$ *and* $\delta > 0$ *such that if* $|w_+ - w_-| \le \delta$ *and* $\|F_0\|_{H^2} \le \varepsilon_0$*, then* (2.5) *has a unique global smooth solution* F *satisfying*

$$\sum_{j=0}^{2} w_{j+1}(t) \|\partial_x^j F(\cdot, t)\|^2 + \int_0^t \sum_{j=1}^{3} w_j(\tau) \|\partial_x^j F(\cdot, \tau)\|^2 \, d\tau \le C,$$

where the weight functions $w_j(t)$ *are given by*

$$w_1(t) = (1+t)^{\frac{1}{2}} (1 + \log(1+t))^{-k}, \ w_j(t) = (1+t)^{j-1} w_1(t)$$

for $j, k > 1$.

Remark 3. (a) The condition (2.6) plays an important role in our proof of Theorem 2.2 (see Lemmas 2.3–2.7, 2.9–2.10 below). This condition enables us to bound the L^1-norm of F for all time. In [7], $s(x) - \bar{s}$ is assumed to be compact support besides the technical condition V; our condition (2.6) is much weaker. In fact, (2.6) asks only some decay properties on $s(x) - \bar{s}$ as $x \to \pm\infty$.

(b) The condition (2.6) can be replaced by the weaker one such as

$$|x|^{\beta} (s(x) - \bar{s}) \in L^1(R) \tag{2.6$'$}$$

for some $\beta > 0$. This is clear following our proof.

(c) In general, we could not bound the L^1-norm of F for all time without the conditions on the decay properties of $s(x) - \bar{s}$ as $x \to \pm\infty$ such as (2.6′). One cannot even bound the total mass of F uniformly in time under the condition $s(x) - \bar{s} \in L^1$. From this point of view, (2.6′) is optimal.

The local existence and uniqueness of the solution to (2.5) in $X(0, T)$ is standard, so to get the global existence, we will prove uniform estimates on the solution of (2.5). Hence, from now on, we assume the local existence in $X(0, T)$ for some $T > 0$.

The following L^1-estimate follows from the standard contraction property of the porous media type equation and will play a fundamental role in the rest of this section.

LEMMA 2.3. *Under the conditions of Theorem* 2.2*, as long as the solution exists in* $X(0, T)$*, there exist positive constants* C_1 *and* C_2*, such that*

$$\|\phi(\cdot, t)\|_{L^1} \le C_1 \|F(\cdot, t)\|_{L^1} \le C_2 (\|\phi_0\|_{L^1} + \delta). \tag{2.7}$$

Proof. We present here a formal argument which can easily be made rigorous by using any sequence approximating the sign function and passing into the limit by means of the *Lebesgue dominated convergence theorem*. Observe that $h = sign(\phi) = sign(F)$. Let us multiply the equation in (2.4) by $a^{-1}h$, then by integrating over

$[0,t]\times(-\infty,+\infty)$, it follows that

$$\begin{aligned}&\int_{-\infty}^{+\infty}a^{-1}|\phi|(x,t)dx+\int_0^t\int_{-\infty}^{+\infty}sign'(F)F_x^2dxd\tau\\&\le C\int_{-\infty}^{+\infty}a^{-1}|\phi_0|(x)dx+C\left|\int_0^t\int_{-\infty}^{+\infty}(a-a_1)\tilde{w}_tsign(F)dxd\tau\right|\\&+\left|\int_0^t\int_{-\infty}^{+\infty}(fF^2)_xF_xsign'(F)dxd\tau\right|\\&\le C(\|\phi_0\|_{L^1}+\delta).\end{aligned}\tag{2.8}$$

Here, we have used the following facts:

$$\begin{aligned}&\left|\int_0^t\int_{-\infty}^{+\infty}(a-a_1)\tilde{w}_tsign(F)dxd\tau\right|\\&\le C\int_0^t\int_{-\infty}^{+\infty}|s-\bar{s}||\tilde{w}_t|dxd\tau\\&\le C\int_0^t\int_{-\infty}^{+\infty}(1+t)^{-\frac{3}{2}}|x(s-\bar{s})||\tilde{w}'(\eta)|dxd\tau\\&\le C\delta,\end{aligned}\tag{2.9}$$

$$\begin{aligned}&\int_0^t\int_{-\infty}^{+\infty}(fF^2)_xF_xsign'(F)dxd\tau\\&=\int_0^t\int_{-\infty}^{+\infty}F_x(2fF_x+f_FFF_x+f_{\tilde{w}}F\tilde{w}_x)F\delta_{\{F=0\}}dxd\tau\\&=0.\end{aligned}\tag{2.10}$$

Hence (2.8) gives the proof of (2.7). □

With the help of Lemma 2.3, we can make the energy estimates on F.

Lemma 2.4. *Under the hypotheses of Theorem 2.2, there exists $\varepsilon_*>0$ such that if $0<\varepsilon<\varepsilon_*$ and $N(T)\le\varepsilon$, then we have*

$$\|F(\cdot,t)\|^2+\int_0^t\|F_x(\cdot,\tau)\|^2d\tau\le C(\|F_0\|^2+\delta)\tag{2.11}$$

for $0\le t\le T$.

Proof. Let us multiply (2.4) by $a^{-1}F$ and integrate the result over $[0,t]\times(-\infty,+\infty)$; we then get

$$\begin{aligned}&\int_{-\infty}^{+\infty}\frac{1}{2}a^{-1}F\phi(x,t)dx+\int_0^t\int_{-\infty}^{+\infty}F_x^2dxd\tau\\&\le\int_{-\infty}^{+\infty}\frac{1}{2}a^{-1}F_0\phi_0dx+\left|\int_0^t\int_{-\infty}^{+\infty}a^{-1}(a-a_1)(\tilde{w}^{-\gamma})_{xx}Fdxd\tau\right|\\&+\left|\int_0^t\int_{-\infty}^{+\infty}\frac{1}{2}a^{-1}\psi_2(\tilde{w})F^2\tilde{w}_tdxd\tau\right|+\left|\int_0^t\int_{-\infty}^{+\infty}(fF^2)_xF_xdxd\tau\right|\\&\equiv\int_{-\infty}^{+\infty}\frac{1}{2}a^{-1}F_0\phi_0dx+I_1+I_2+I_3,\end{aligned}\tag{2.12}$$

with $\psi_2(\tilde{w})F^2 = \phi^2\psi'(\tilde{w})$.

We use I_1, I_2, and I_3 step-by-step as follows:

$$\begin{aligned} I_1 &= \left| \int_0^t \int_{-\infty}^{+\infty} a^{-1}(a-a_1)(\tilde{w})^{-\gamma}_{xx} F dx d\tau \right| \\ &\le C\delta\varepsilon \int_0^t (1+\tau)^{-\frac{3}{2}} \|x(s-\bar{s})\|_{L^1}\ d\tau \\ &\le C\delta\varepsilon, \end{aligned} \tag{2.13}$$

$$\begin{aligned} I_2 &= \left| \int_0^t \int_{-\infty}^{+\infty} \frac{1}{2} a^{-1}\psi_2(\tilde{w})F^2 \tilde{w}_t dx d\tau \right| \\ &\le C \int_0^t \|F\|_{L^\infty} \|\tilde{w}_t\|_{L^\infty} \|F\|_{L^1} dx d\tau \\ &\le C\delta \int_0^t \|F\|^{\frac{1}{2}} \|F_x\|^{\frac{1}{2}} (1+\tau)^{-1} d\tau \\ &\le C\delta \left(\int_0^t \|F\|^2 \|F_x\|^2 d\tau + \int_0^t (1+\tau)^{-\frac{4}{3}} d\tau \right) \\ &\le C\delta \left(1 + \varepsilon^2 \int_0^t \|F_x\|^2\ d\tau \right), \end{aligned} \tag{2.14}$$

$$\begin{aligned} I_3 &= \left| \int_0^t \int_{-\infty}^{+\infty} (fF^2)_x F_x dx d\tau \right| \\ &\le \left(\frac{1}{2} + C\varepsilon \right) \int_0^t \|F_x\|^2\ d\tau + C\delta^2 \int_0^t \|F\|^4_{L^\infty}\ d\tau \\ &\le \left(\frac{1}{2} + C\varepsilon \right) \int_0^t \|F_x\|^2\ d\tau + C\delta^2 \int_0^t \|F\|^2 \|F_x\|^2\ d\tau \\ &\le \left(\frac{1}{2} + C\varepsilon \right) \int_0^t \|F_x\|^2\ d\tau. \end{aligned} \tag{2.15}$$

Due to the smallness of δ and ε, we conclude from (2.12)–(2.15) that

$$\|F(\cdot,t)\|^2 + \int_0^t \|F_x(\cdot,\tau)\|^2 d\tau \le C(\|F_0\|^2 + \delta), \tag{2.16}$$

which completes the proof of Lemma 2.4. □

For higher order estimates, we use the problem (2.5) to obtain the following results.

LEMMA 2.5. *Under the same conditions of Lemma* 2.4, *we have*

$$\|F_x(\cdot,t)\|^2 + \int_0^t \|F_{xx}(\cdot,\tau)\|^2\ d\tau \le C(\|F_0\|^2_{H^1} + \delta). \tag{2.17}$$

Proof. Let us multiply the equation in (2.5) by F_{xx}, then

$$\begin{aligned}&\int_{-\infty}^{+\infty} F_x^2(x,t)dx + \int_0^t \int_{-\infty}^{+\infty} F_{xx}^2(x,\tau)\ dxd\tau \\ &\le C\left(\|F_{0x}\|^2 + \left|\int_0^t \int_{-\infty}^{+\infty} \tilde{w}_t F_{xx}\ dxd\tau\right|\right. \\ &\qquad \left. + \left|\int_0^t \int_{-\infty}^{+\infty} (fF^2)_{xx} F_{xx}\ dxd\tau\right|\right),\end{aligned} \tag{2.18}$$

which implies, with the help of the Cauchy–Schwarz inequality and Lemma 2.1, that

$$\begin{aligned}&\int_{-\infty}^{+\infty} F_x^2(x,t)dx + \int_0^t \int_{-\infty}^{+\infty} F_{xx}^2(x,\tau)\ dxd\tau \\ &\le C(\|F_{0x}\|^2 + \delta^2) + C\int_0^t \int_{-\infty}^{+\infty} (fF^2)_{xx}^2\ dxd\tau.\end{aligned} \tag{2.19}$$

We bound the last term in (2.19) as follows:

$$\begin{aligned}&\int_0^t \int_{-\infty}^{+\infty} (fF^2)_{xx}^2\ dxd\tau \\ &\le C\int_0^t \int_{-\infty}^{+\infty} [(|F| + |F_x| + |w_x|)^2 F_x^2 + F^2 F_{xx}^2 + F^4(\tilde{w}_{xx}^2 + \tilde{w}_x^4)]\ dxd\tau \\ &\le C\varepsilon^2\delta^2 + C\varepsilon \int_0^t \int_{-\infty}^{+\infty} F_{xx}^2(\tau, x)\ dxd\tau.\end{aligned} \tag{2.20}$$

Then, by (2.19)–(2.20) and the estimates in Lemma 2.4, we get (2.17). □

We now turn to the third order estimates. For this purpose, we differentiate the equation in (2.5) with respect to x

$$\begin{aligned}&F_{tx} + a\psi(\tilde{w})F_{xxx} + (a\psi(\tilde{w}))_x F_{xx} - (\psi(\tilde{w})(a - a_1)(\tilde{w}^{-\gamma})_{xx})_x \\ &\quad + (\psi_1(\tilde{w})F\tilde{w}_t)_x - (a\psi(\tilde{w})(fF^2)_{xx})_x = 0.\end{aligned} \tag{2.21}$$

Multiplying (2.21) by F_{xxx} and then integrating it over $[0,t] \times (-\infty, +\infty)$, one has

$$\begin{aligned}&\int_{-\infty}^{+\infty} F_{xx}^2(\cdot,t)dx + \int_0^t \int_{-\infty}^{+\infty} F_{xxx}^2(\tau,x)\ dxd\tau \\ &\le C\left(\|F_{0xx}\|^2 + \int_0^t \int_{-\infty}^{+\infty} ((a\psi)_x F_{xx})^2\ dxd\tau + \int_0^t \int_{-\infty}^{+\infty} (\psi_1 \tilde{w} F \tilde{w}_t)_x^2\ dxd\tau \right. \\ &\qquad \left. + \int_0^t \int_{-\infty}^{+\infty} [((\psi(\tilde{w})(a - a_1)(\tilde{w}^{-\gamma})_{xx})_x^2 + (a\psi(\tilde{w})(fF^2)_{xx})_x^2]\ dxd\tau \right) \\ &\le C(\|F_0\|_{H^2}^2 + \delta^2) + C\varepsilon \int_0^t \int_{-\infty}^{+\infty} F_{xxx}^2(\tau,x)\ dxd\tau,\end{aligned} \tag{2.22}$$

which can be summarized as follows.

LEMMA 2.6. *Under the same conditions as Lemma 2.4, one has*

$$\|F_{xx}(\cdot,t)\|^2 + \int_0^t \|F_{xxx}(\cdot,\tau)\|^2\ d\tau \le C(\|F_0\|_{H^2}^2 + \delta). \tag{2.23}$$

From Lemma 2.4–2.6, we can conclude the following.

Lemma 2.7. *Under the same conditions as Theorem* 2.2, *there exists* $\varepsilon_* > 0$ *such that if* $0 < \varepsilon < \varepsilon_*$ *and* $N(T) \leq \varepsilon$, *then it follows that*

$$\|F(t)\|_{H^2}^2 + \int_0^t (\|F_t\|_{H^1}^2 + \|F_x\|_{H^2}^2)(\tau)\ d\tau \leq C_0(\|F_0\|_{H^2}^2 + \delta)$$

for all $0 \leq t \leq T$, *where* C_0 *is a positive constant independent of* t.

With the help of the previous lemmas we obtain the global existence of the solution $F(x,t)$ to (2.5).

Theorem 2.8. *Under the same conditions as Theorem* 2.2, (2.5) *has a unique global smooth solution* $F(x,t)$ *which tends to zero uniformly in* H^1 *as* t *goes to infinity.*

Proof. We choose δ, ε_0, and ε small such that $C_0(\varepsilon_0^2 + \delta) \leq \varepsilon^3$, $\varepsilon \leq \frac{1}{4}$ so that all the previous arguments are valid. Then, due to the local results, there exists a positive t_1 such that the solution $F(x,t)$ exists in $(-\infty, +\infty) \times [0, t_1]$ and satisfies

$$N(t)^2 \leq 4N^2(0) \text{ for all } t \in [0, t_1].$$

We can apply the L^1-estimate for F of Lemma 2.3 and then Lemma 2.7 in $0 \leq t \leq t_1$. Therefore, it follows that

$$N^2(t) \leq C_0(\|F_0\|_{H^2}^2 + \delta) \leq \varepsilon^3 \text{ for all } t \in [0, t_1].$$

By iterating the above procedure, a standard continuity argument allows us to establish the global existence in time for the solution to (2.5).

Now, from Lemma 2.7 and the above argument, we have

$$\|F(t)\|_{H^2}^2 + \int_0^t (\|F_x\|_{H^2}^2 + \|F_t\|_{H^1}^2)(\tau)\ d\tau \leq C_0(\|F_0\|_{H^2}^2 + \delta) \ \text{ for all } t > 0. \tag{2.24}$$

From (2.24) we know that

$$\|F_x(t)\|^2 + \int_0^{+\infty} \left| \frac{d}{dt} \|F_x(t)\|^2 \right|\ dt \leq C,$$

which implies

$$\lim_{t \to +\infty} \|F_x(t)\|^2 = 0.$$

Then, the Sobolev inequality implies

$$\begin{aligned} \lim_{t \to +\infty} \|F(t)\|^2 &\leq \lim_{t \to +\infty} \|F(t)\|_{L^\infty} \|F(t)\|_{L^1} \\ &= O(1) \lim_{t \to +\infty} \|F_x(t)\|^{\frac{1}{2}} \\ &= 0, \end{aligned}$$

which completes the proof of this theorem. □

By using the weighted energy method, we can prove the following decay estimates.

Lemma 2.9. *Let* F *be the solution to* (2.5) *obtained in Theorem* 2.8, *then*

$$\begin{aligned} &w_1(t)\|F(t)\|^2 + w_2(t)\|F_x(t)\|^2 \\ &+ \int_0^t (w_1(\tau)\|F_x(\tau)\|^2 + w_2(\tau)\|F_{xx}(\tau)\|^2)\ d\tau \leq C. \end{aligned} \tag{2.25}$$

Proof. Let us multiply (2.5) by $w_2(t)F_{xx}$, then we get

$$\begin{aligned}&\left(\frac{1}{2}w_2(t)F_x^2\right)_t - a\psi(\tilde{w})w_2(t)F_{xx}^2 - \frac{1}{2}w_2'(t)F_x^2 - \psi_1(\tilde{w})FF_{xx}\tilde{w}_t w_2(t)\\ &= -w_2(t)\psi(\tilde{w})(a-a_1)(\tilde{w}^{-\gamma})_{xx}F_{xx} - a\psi(\tilde{w})(fF^2)_{xx}F_{xx}w_2(t) + (\cdots)_x,\end{aligned}\tag{2.26}$$

where $(\cdots)_x$ denotes the term which does not need to be computed explicitly since it will disappear by integrating in x. Then one has

$$\begin{aligned}&w_2(t)\|F_x(\cdot,t)\|^2 + \int_0^t w_2(\tau)\|F_{xx}(\cdot,\tau)\|^2\ d\tau\\ &\le C_1\left(\|F_{0x}\|^2 + \left|\int_0^t\int_{-\infty}^{+\infty} w_2'(\tau)F_x^2 dxd\tau\right|\right.\\ &\qquad + \left|\int_0^t\int_{-\infty}^{+\infty} \tilde{w}_t^2 w_2(\tau)(a-a_1)^2\ dxd\tau\right|\\ &\qquad \left.+ \left|\int_0^t\int_{-\infty}^{+\infty} F^2\tilde{w}_t^2 w_2(\tau)\ dxd\tau\right| + \int_0^t\int_{-\infty}^{+\infty} w_2(\tau)(fF^2)_{xx}^2\ dxd\tau\ \right)\\ &\equiv C_1(\|F_{0x}\|^2 + J_1 + J_2 + J_3 + J_4).\end{aligned}\tag{2.27}$$

On the other hand, if we multiply (2.4) by $a^{-1}w_1(t)F$, we get

$$\begin{aligned}&\left(\frac{1}{2}F\phi a^{-1}w_1(t)\right)_t + w_1(t)F_x^2 - \frac{1}{2}w_1'(t)a^{-1}\psi_1(\tilde{w})F^2\\ &= \frac{1}{2}a^{-1}w_1(t)F^2\tilde{w}_t - a^{-1}w_1(t)(a-a_1)F(\tilde{w}^{-\gamma})_{xx}\\ &+ w_1(t)F_x(fF^2)_x + (\cdots)_x,\end{aligned}\tag{2.28}$$

which, integrated on $[0,t]\times(-\infty,+\infty)$, yields

$$\begin{aligned}&w_1(t)\|F(\cdot,t)\|^2 + \int_0^t w_1(t)\|F_x(\tau)\|^2 d\tau\\ &\le C_2\left(\|F_0\|^2 + \left|\int_0^t\int_{-\infty}^{+\infty} w_1'(\tau)F^2\ dxd\tau\right|\right.\\ &\qquad + \left|\int_0^t\int_{-\infty}^{+\infty} w_1(\tau)F^2\tilde{w}_t\ dxd\tau\right| + \left|\int_0^t\int_{-\infty}^{+\infty} w_1(\tau)\tilde{w}_t F(a-a_1)\ dxd\tau\right|\\ &\qquad \left.+ \left|\int_0^t\int_{-\infty}^{+\infty} w_1(\tau)F_x(fF^2)_x\ dxd\tau\right|\ \right)\\ &\equiv C_2(\|F_0\|^2 + J_5 + J_6 + J_7 + J_8).\end{aligned}\tag{2.29}$$

By calculating $K\times(2.29)+(2.27)$ with a $K>0$ to be determined later, we have

$$\begin{aligned}&Kw_1(t)\|F(t)\|^2 + w_2(t)\|F_x(t)\|^2\\ &+ \int_0^t (Kw_1(\tau)\|F_x(\tau)\|^2 + w_2(\tau)\|F_{xx}(\tau)\|^2)\ d\tau\\ &\le (C_1\|F_{0x}\|^2 + KC_2\|F_0\|^2)\\ &+ C_1(J_1+J_2+J_3+J_4) + KC_2(J_5+J_6+J_7+J_8).\end{aligned}\tag{2.30}$$

The following inequalities will be used to estimate the terms on $J_i (i = 1, \dots, 8)$:

$$\begin{aligned} |w_1'(t)| &= \left| \frac{1}{2}(1+t)^{-1}w_1(t) - k(1+\log(1+t))^{-1}(1+t)^{-1}w_1(t) \right| \\ &\le C(1+t)^{-1}w_1(t), \end{aligned} \tag{2.31}$$

$$|w_2'(t)| \le C(1+t)^{-1}w_2(t) = Cw_1(t). \tag{2.32}$$

Then, by choosing K large enough, one has

$$C_1 J_1 \le \frac{1}{2}K \int_0^t w_1(\tau)\|F_x(\tau)\|^2 \, d\tau, \tag{2.33}$$

$$J_2 \le C\delta^2 \int_0^t (1+\tau)^{-3} w_2(\tau) \, d\tau \le C\delta^2. \tag{2.34}$$

To estimate J_3, observe that the following inequality on F holds:

$$\|F\|_{L^\infty} \le C\|F_x\|^{\frac{2}{3}} \tag{2.35}$$

since

$$\begin{aligned} \|F\|_{L^\infty} &\le C\|F\|^{\frac{1}{2}}\|F_x\|^{\frac{1}{2}} \\ &\le C\|F\|_{L^\infty}^{\frac{1}{4}}\|F_x\|^{\frac{1}{2}}\|F\|_{L^1}^{\frac{1}{4}}. \end{aligned}$$

Then we see that

$$\begin{aligned} C_1 J_3 &= C_1 \left| \int_0^t \int_{-\infty}^{+\infty} F^2 \tilde{w}_t^2 w_2(\tau) \, dx d\tau \right| \\ &\le C \int_0^t w_2(\tau)\|\tilde{w}_t\|_{L^\infty}\|F\|_{L^\infty}\|F\|_{L^1} \, d\tau \\ &\le C\delta^2 \int_0^t (1+\tau)^{-2} w_2(\tau)\|F_x\|^{\frac{2}{3}} d\tau \\ &\le C\delta^2 + \frac{1}{4}K \int_0^t w_1(\tau)\|F_x(\cdot,\tau)\|^2 \, d\tau. \end{aligned} \tag{2.36}$$

Similarly, we can estimate J_5, J_6, and J_7 as follows:

$$\begin{aligned} C_2 K J_5 &= C_2 K \left| \int_0^t \int_{-\infty}^{+\infty} w_1'(\tau) F^2 \, dx d\tau \right| \\ &\le CK \int_0^t (1+\tau)^{-1} w_1(\tau)\|F_x\|^{\frac{2}{3}} d\tau \\ &\le CK + \frac{1}{8}K \int_0^t w_1(\tau)\|F_x(\cdot,\tau)\|^2 \, d\tau, \end{aligned} \tag{2.37}$$

$$\begin{aligned} C_2 K J_6 &= C_2 K \left| \int_0^t \int_{-\infty}^{+\infty} w_1(\tau) F^2 \tilde{w}_t \; dx d\tau \right| \\ &\leq CK\delta + \frac{1}{16} K \int_0^t w_1(\tau) \|F_x(\cdot,\tau)\|^2 \; d\tau, \end{aligned} \tag{2.38}$$

$$\begin{aligned} C_2 K J_7 &= C_2 K \left| \int_0^t \int_{-\infty}^{+\infty} w_1(\tau) \tilde{w}_t F(a-a_1) \; dx d\tau \right| \\ &\leq CK\delta + \frac{1}{32} K \int_0^t w_1(\tau) \|F_x(\cdot,\tau)\|^2 \; d\tau. \end{aligned} \tag{2.39}$$

To estimate J_8, we have

$$\begin{aligned} & C_2 K J_8 \\ &= C_2 K \left| \int_0^t \int_{-\infty}^{+\infty} w_1(\tau) F_x (fF^2)_x \; dx d\tau \right| \\ &\leq C \left| \int_0^t \int_{-\infty}^{+\infty} w_1(\tau) F_x (2fFF_x + f_F F^2 F_x + f_{\tilde{w}} \tilde{w}_x F^2) \; dx d\tau \right| \\ &\leq C\varepsilon K \int_0^t w_1(\tau) \|F_x(\cdot,\tau)\|^2 d\tau + CK \int_0^t \int_{-\infty}^{+\infty} w_1(\tau) \tilde{w}_x^2 F^2 \; dx d\tau \\ &\leq CK(\delta+\varepsilon) \int_0^t w_1(\tau) \|F_x(\cdot,\tau)\|^2 d\tau + CK\delta. \end{aligned} \tag{2.40}$$

We now deal with the term J_4. Noting that

$$\begin{aligned} (fF^2)_{xx} &= (2FF_x f + f_F F^2 F_x + f_{\tilde{w}} \tilde{w}_x F^2)_x \\ &= (2fF + f_F F^2) F_{xx} + (2f + 4f_F F + f_{FF} F^2) F_x^2 \\ &\quad + (4 f_{\tilde{w}} F + 2 f_{F\tilde{w}} F^2) F_x \tilde{w}_x + (f_{\tilde{w}} \tilde{w}_{xx} + f_{\tilde{w}\tilde{w}} \tilde{w}_x^2) F^2, \end{aligned}$$

we have

$$\begin{aligned} J_4 &= \left| \int_0^t \int_{-\infty}^{+\infty} w_2(\tau) (fF^2)_{xx}^2 \; dx d\tau \right| \\ &\leq C\varepsilon \int_0^t w_2(\tau) \|F_{xx}(\cdot,\tau)\|^2 \; d\tau + C \int_0^t \int_{-\infty}^{+\infty} F_x^4 w_2(\tau) \; dx d\tau \\ &\quad + C \left| \int_0^t \int_{-\infty}^{+\infty} w_2(\tau) F_x^2 \tilde{w}_x^2 F \; dx d\tau \right| \\ &\quad + C \int_0^t \int_{-\infty}^{+\infty} w_2(\tau) F^4 (\tilde{w}_{xx}^2 + \tilde{w}_x^4) \; dx d\tau \\ &\equiv C\varepsilon \int_0^t w_2(\tau) \|F_{xx}(\cdot,\tau)\|^2 \; d\tau + C(J_9 + J_{10} + J_{11}). \end{aligned} \tag{2.41}$$

We can estimate J_{10} and J_{11} in the following way:

$$\begin{aligned} J_{10} &= \left| \int_0^t \int_{-\infty}^{+\infty} w_2(\tau) F_x^2 \tilde{w}_x^2 F \ dx d\tau \right| \\ &\leq C\varepsilon\delta^2 \int_0^t w_1(\tau) \|F_x(\cdot,\tau)\|^2 \ d\tau, \end{aligned} \tag{2.42}$$

$$\begin{aligned} J_{11} &= \int_0^t \int_{-\infty}^{+\infty} w_2(\tau) F^4 (\tilde{w}_{xx}^2 + \tilde{w}_x^4) \ dx d\tau \\ &\leq C\delta^2\varepsilon^2 \int_0^t w_1(\tau) \|F_x(\cdot,\tau)\|^2 \ d\tau, \end{aligned} \tag{2.43}$$

while J_9 can be bounded as follows:

$$\begin{aligned} J_9 &= \int_0^t \int_{-\infty}^{+\infty} F_x^4 w_2(\tau) \ dx d\tau \\ &\leq C\varepsilon^2 \int_0^t w_2(\tau) \|F_{xx}\|^2 \ d\tau + C \int_0^t w_2(\tau) \|F_x(\tau)\|^2 \|F_x(\tau)\|^2 \ d\tau. \end{aligned} \tag{2.44}$$

Due to the smallness of δ and ε, by choosing K large enough, we deduce

$$\begin{aligned} & w_1(t)\|F(t)\|^2 + w_2(t)\|F_x(t)\|^2 \\ & + \int_0^t (w_1(\tau)\|F_x(\tau)\|^2 + w_2(\tau)\|F_{xx}(\tau)\|^2) \ d\tau \\ & \leq C \left(1 + \int_0^t w_2(\tau)\|F_x(\tau)\|^2 \|F_x(\tau)\|^2 \ d\tau \right). \end{aligned} \tag{2.45}$$

Therefore, from (2.45), it follows that

$$w_2(t)\|F_x(t)\|^2 \leq C \left(1 + \int_0^t w_2(\tau)\|F_x(\tau)\|^2 \|F_x(\tau)\|^2 \ d\tau \right),$$

which implies, with the help of Gronwall inequality, that

$$w_2(t)\|F_x(t)\|^2 \leq C$$

and

$$\int_0^t w_2(\tau)\|F_x(\tau)\|^2 \|F_x(\tau)\|^2 \ d\tau \leq C. \tag{2.46}$$

Hence, (2.45) and (2.46) give the proof of this lemma. □

The following lemma contains the decay rates for the derivatives of F, which will be useful in the next section.

Lemma 2.10. *The solution F to* (2.5)*, obtained in Theorem* 2.8*, satisfies*

$$\begin{aligned} w_3(t)\|F_t(t)\|^2 + w_4(t)\|F_{tx}\|^2 + \int_0^t (w_3(\tau)\|F_{tx}(\tau)\|^2 + w_4(\tau)\|F_{txx}(\tau)\|^2) \ d\tau & \\ \leq C\delta, & \end{aligned} \tag{2.47}$$

$$\|F_t\|_{L^\infty} \leq C w_3(t)^{-\frac{1}{4}} w_4(t)^{-\frac{1}{4}}.$$

Proof. It is sufficient to prove (2.47), since the estimate for $\|F_t\|_{L^\infty}$ can be derived from (2.47) by using Sobolev inequality.

Let us differentiate $(2.5)_1$ in t, then

$$\begin{aligned}&F_{tt} + a\psi(\tilde{w})F_{txx} + a\psi'(\tilde{w})\tilde{w}_t F_{xx} - [\psi(\tilde{w})(a-a_1)(\tilde{w}^{-\gamma})_{xx}]_t\\&-(\psi_1(\tilde{w})F\tilde{w}_t)_t - [a\psi(\tilde{w})(fF^2)_{xx}]_t = 0.\end{aligned} \tag{2.48}$$

If we multiply (2.48) by $a^{-1}w_3(t)F_t$, we get

$$\begin{aligned}&\left(\frac{1}{2}a^{-1}w_3(t)F_t^2\right)_t - \psi(\tilde{w})w_3(t)F_{tx}^2 + \frac{1}{2}F_t^2\psi(\tilde{w})_{xx}w_3(t) - \frac{1}{2}F_t^2a^{-1}w_3'(t)\\&+\psi'(\tilde{w})\tilde{w}_tF_{xx}w_3(t)F_t - a^{-1}[\psi(\tilde{w})(a-a_1)(\tilde{w}^{-\gamma})_{xx}]_tw_3(t)F_t\\&-a^{-1}(\psi_1(\tilde{w})F\tilde{w}_t)_tw_3(t)F_t - [\psi(\tilde{w})(fF^2)_{xx}]_tw_3(t)F_t + \{\cdots\}_x = 0.\end{aligned} \tag{2.49}$$

From the proof of Lemma 2.9 and $(2.5)_1$ it is clear that

$$\begin{aligned}&\int_0^t w_2(\tau)\|F_t(\cdot,\tau)\|^2\ d\tau\\&\le C\left(\int_0^t w_2(t)\|F_{xx}(\tau)\|^2 d\tau + \int_0^t\int_{-\infty}^{+\infty}(a-a_1)^2\tilde{w}_t^2w_2(\tau)\ dxd\tau\right.\\&\qquad\left.+\int_0^t\int_{-\infty}^{+\infty}F^2\tilde{w}_t^2w_2(\tau)\ dxd\tau + \int_0^t\int_{-\infty}^{+\infty}(fF^2)_{xx}^2w_2(\tau)\ dxd\tau\right)\\&\le C.\end{aligned} \tag{2.50}$$

Moreover, we have

$$a^{-1}(\psi_1(\tilde{w})F\tilde{w}_t)_tw_3(t)F_t = O(1)[\tilde{w}_tw_3(t)F_t^2 + (\tilde{w}_t^2+\tilde{w}_{tt})w_3(t)FF_t],$$

$$a^{-1}[\psi(\tilde{w})(a-a_1)(\tilde{w}^{-\gamma})_{xx}]_tw_3(t)F_t = O(1)(a-a_1)(\tilde{w}_t^2+\tilde{w}_{tt})w_3(t)F_t,$$

$$[\psi(\tilde{w})(fF^2)_{xx}]_tw_3(t)F_t = O(1)\tilde{w}_t(fF^2)_{xx}w_3(t)F_t - \psi(\tilde{w})(fF^2)_{xxt}w_3(t)F_t.$$

Now, integrating (2.49) and integrating by parts, we have

$$\begin{aligned}&w_3(t)\|F_t(t)\|^2 + \int_0^t w_3(\tau)\|F_{tx}(\tau)\|^2\ d\tau\\&\le C + C\int_0^t\int_{-\infty}^{+\infty}F_t^2w_2(\tau)\ dxd\tau + \int_0^t\int_{-\infty}^{+\infty}\tilde{w}_tw_3(\tau)F_{xx}^2\ dxd\tau\\&\quad+\int_0^t\int_{-\infty}^{+\infty}[(a-a_1)^2+F^2](\tilde{w}_t^2+\tilde{w}_{tt})^2w_3(\tau)(1+\tau)\ dxd\tau\\&\quad+\int_0^t\int_{-\infty}^{+\infty}w_2(\tau)(fF^2)_{xx}^2\ dxd\tau + \left|\int_0^t\int_{-\infty}^{+\infty}\psi(\tilde{w})(fF^2)_{xxt}w_3(\tau)F_t\ dxd\tau\right|\\&\le C\left(1+\left|\int_0^t\int_{-\infty}^{+\infty}w_3(\tau)(fF^2)_{xt}(F_{tx}+F_t\tilde{w}_x)\ dxd\tau\right|\right).\end{aligned} \tag{2.51}$$

To bound the previous terms, we observe that

$$\begin{aligned}&\left|\int_0^t \int_{-\infty}^{+\infty} w_3(\tau)(fF^2)_{xt} F_t \tilde{w}_x \ dxd\tau\right| \\ &\le C + C\int_0^t \int_{-\infty}^{+\infty} w_3(\tau)(fF^2)^2_{xt} \ dxd\tau\end{aligned}$$

and

$$\begin{aligned}&\left|\int_0^t \int_{-\infty}^{+\infty} w_3(\tau)(fF^2)_{xt} F_{tx} \ dxd\tau\right| \\ &\le \varepsilon_3 \int_0^t \int_{-\infty}^{+\infty} w_3(\tau)F^2_{tx} \ dxd\tau + C(\varepsilon_3)\int_0^t \int_{-\infty}^{+\infty} w_3(\tau)(fF^2)^2_{xt} \ dxd\tau.\end{aligned}$$

Then choosing ε_3 sufficient small, we have

$$\begin{aligned}&w_3(t)\|F_t(t)\|^2 + \int_0^t w_3(\tau)\|F_{tx}(\tau)\|^2 \ d\tau \\ &\le C\left(1 + \int_0^t \int_{-\infty}^{+\infty} w_3(\tau)(fF^2)^2_{xt} \ dxd\tau\right).\end{aligned} \tag{2.52}$$

Moreover, since

$$\begin{aligned}(fF^2)^2_{xt} &= [2fFF_x, f_F F^2 F_x + f_{\tilde{w}} \tilde{w}_x F^2]^2_t \\ &= O(1)(|F|F^2_{xt} + F^2_t F^2_x + \tilde{w}^2_t F^2_x F^2 \\ &\quad + F^2 F^2_t \tilde{w}^2_x + (\tilde{w}^2_{tx} + \tilde{w}^2_t \tilde{w}^2_x)F^4),\end{aligned}$$

it follows that

$$\begin{aligned}&\int_0^t \int_{-\infty}^{+\infty} w_3(\tau)(fF^2)^2_{xt} \ dxd\tau) \\ &\le C\left(\varepsilon \int_0^t w_3(\tau)\|F_{tx}\|^2 d\tau + A_1 + A_2 + A_3 + A_4\right),\end{aligned} \tag{2.53}$$

where, as in the previous estimates we get

$$A_1 = \int_0^t \int_{-\infty}^{+\infty} w_3(\tau) F^2_t F^2_x \ dxd\tau \le C, \tag{2.54}$$

$$A_2 = \int_0^t \int_{-\infty}^{+\infty} w_3(\tau) \tilde{w}^2_t F^2_x F^2 \ dxd\tau \le C\delta^2\varepsilon^2, \tag{2.55}$$

$$A_3 = \int_0^t \int_{-\infty}^{+\infty} w_3(\tau) F^2 F^2_t \tilde{w}^2_x \ dxd\tau \le C\delta^2\varepsilon^2, \tag{2.56}$$

and

$$A_4 = \int_0^t \int_{-\infty}^{+\infty} w_3(\tau)(\tilde{w}^2_{tx} + \tilde{w}^2_t \tilde{w}^2_x)F^4 \ dxd\tau \le C\delta^2. \tag{2.57}$$

Thus, we conclude from (2.52)—(2.57) that

$$w_3(t)\|F_t(t)\|^2 + \int_0^t w_3(\tau)\|F_{tx}(\tau)\|^2 \, d\tau \le C, \tag{2.58}$$

which completes the proof of the first part of (2.47).

Let us turn to the second part of (2.47). For this purpose, we multiply (2.48) by $w_4(t)F_{txx}$. After similar calculations as before, by virtue of (2.58), we get

$$\begin{aligned} & w_4(t)\|F_{tx}(t)\|^2 + \int_0^t w_4(\tau)\|F_{txx}(\tau)\|^2 \, d\tau \\ & \le C(1 + B_1 + B_2 + B_3 + B_4), \end{aligned} \tag{2.59}$$

where

$$B_1 = \int_0^t \int_{-\infty}^{+\infty} \tilde{w}_t^2 w_4(\tau) F_{xx}^2 \, dx d\tau \le C\delta^2,$$

$$B_2 = \int_0^t \int_{-\infty}^{+\infty} (\tilde{w}_t^2 + \tilde{w}_{tt})^2 w_4(\tau)(a - a_1)^2 \, dx d\tau \le C\delta^2,$$

$$B_3 = \int_0^t \int_{-\infty}^{+\infty} w_4(\tau)[\psi_1(\tilde{w}) F \tilde{w}_t]_t^2 \, dx d\tau \le C\delta^2,$$

and

$$\begin{aligned} B_4 &= \int_0^t \int_{-\infty}^{+\infty} w_4(\tau)[\psi(\tilde{w}(fF^2)_{xx}]_t^2 \, dx d\tau \\ & \le C\delta^2 + C\int_0^t \int_{-\infty}^{+\infty} w_4(\tau)(fF^2)_{txx}^2 \, dx d\tau. \end{aligned}$$

Hence, one has

$$\begin{aligned} & w_4(t)\|F_{tx}(t)\|^2 + \int_0^t w_4(\tau)\|F_{txx}(\tau)\|^2 \, d\tau \\ & \le C\left(1 + \int_0^t \int_{-\infty}^{+\infty} w_4(\tau)(fF^2)_{txx}^2 \, dx d\tau\right). \end{aligned} \tag{2.60}$$

In order to bound the last term in (2.60), we use the following identity:

$$\begin{aligned} & (fF^2)_{txx}^2 \\ & = O(1)[F^2F_{txx}^2 + F_x^2F_{tx}^2 + F^2\tilde{w}_x^2F_{tx}^2 \\ & \qquad + (\tilde{w}_{xx}^2 + \tilde{w}_x^4)F^2F_t^2 + \tilde{w}_{tx}^2F^2F_x^2] \\ & \qquad\qquad + O(1)[(F_t^2F_{xx}^2 + F_{xx}^2F^2\tilde{w}_t^2) + (F_x^4F_t^2 + \tilde{w}_t^2F_x^4) \\ & \qquad\qquad + (F_t^2F_x^2\tilde{w}_x^2 + F^2\tilde{w}_t^2\tilde{w}_x^2F_x^2) \\ & \qquad\qquad + (F^4F_t^2(\tilde{w}_{xx}^2 + \tilde{w}_x^4) + (\tilde{w}_{txx}^2 + \tilde{w}_x^2\tilde{w}_{tx}^2)F^4)] \\ & = [\Gamma_1 + \Gamma_2 + \Gamma_3 + \Gamma_4 + \Gamma_5] + [\Delta_1 + \Delta_2 + \Delta_3 + \Delta_4]. \end{aligned}$$

Then we bound each of these terms. Therefore, we have

$$\begin{aligned}&\int_0^t\int_{-\infty}^{+\infty} w_4(\tau)\Gamma_1\ dxd\tau\\ &\le C\int_0^t\int_{-\infty}^{+\infty} w_4(\tau)F^2F_{txx}^2\ dxd\tau\\ &\le C\varepsilon^2\int_0^t w_4(\tau)\|F_{txx}\|^2\ d\tau,\end{aligned}\tag{2.61}$$

$$\begin{aligned}&\int_0^t\int_{-\infty}^{+\infty} w_4(\tau)\Gamma_2\ dxd\tau\\ &\le C\int_0^t w_4(\tau)\|F_{tx}\|^2\|F_x\|^2\ d\tau + C\varepsilon^2\int_0^t w_4(\tau)\|F_{txx}\|^2 d\tau,\end{aligned}\tag{2.62}$$

$$\int_0^t\int_{-\infty}^{+\infty} w_4(\tau)\Gamma_3\ dxd\tau \le C\varepsilon^2,\tag{2.63}$$

$$\int_0^t\int_{-\infty}^{+\infty} w_4(\tau)\Gamma_4\ dxd\tau \le C\delta^2\varepsilon^2,\tag{2.64}$$

$$\int_0^t\int_{-\infty}^{+\infty} w_4(\tau)\Gamma_5\ dxd\tau \le C\delta^2\varepsilon^2,\tag{2.65}$$

$$\int_0^t\int_{-\infty}^{+\infty} w_4(\tau)\Delta_1\ dxd\tau \le C(1+\delta^2\varepsilon^2),\tag{2.66}$$

$$\int_0^t\int_{-\infty}^{+\infty} w_4(\tau)\Delta_2\ dxd\tau \le C,\tag{2.67}$$

$$\begin{aligned}&\int_0^t\int_{-\infty}^{+\infty} w_4(\tau)\Delta_3\ dxd\tau\\ &\le C\delta^2\int_0^t(\|F_x\|^2+\|F_{xx}\|^2)(w_3(\tau)\|F_t\|^2)\ d\tau + C\delta^2\varepsilon^2\int_0^t w_1(\tau)\|F_x\|^2\ d\tau\\ &\le C,\end{aligned}\tag{2.68}$$

$$\begin{aligned}&\int_0^t\int_{-\infty}^{+\infty} w_4(\tau)\Delta_4\ dxd\tau\\ &\le C\delta^2\int_0^t\|F_t\|^2 w_2(\tau)\ d\tau + C\delta^2\int_0^t w_1(\tau)\|F\|^2\|F_x\|^2\ d\tau\\ &\le C\delta^2.\end{aligned}\tag{2.69}$$

Thus, we see from (2.60)–(2.69) that

$$\begin{aligned} &w_4(t)\|F_{tx}(t)\|^2 + \int_0^t w_4(\tau)\|F_{txx}(\tau)\|^2\ d\tau \\ &\leq C + C\int_0^t w_4(\tau)\|F_{tx}\|^2\|F_x\|^2\ d\tau. \end{aligned} \tag{2.70}$$

Then the Gronwall inequality implies that

$$w_4(t)\|F_{tx}(t)\|^2 + \int_0^t w_4(\tau)\|F_{txx}(\tau)\|^2\ d\tau \leq C. \tag{2.71}$$

Equation (2.47) follows from (2.58) and (2.71). □

COROLLARY 2.11. *The solution F to* (2.5)*, obtained in Theorem* 2.8*, satisfies*

$$w_3(t)\|F_{xx}\|^2 \leq C,\ \|F_{xx}\|_{L^\infty} \leq C(w_3(t)w_4(t))^{-\frac{1}{4}},$$

$$\|F_x\|^2_{L^\infty} \leq Cw_3(t)^{-\frac{1}{2}}w_2(t)^{-\frac{1}{2}}.$$

Proof. From (2.5), we see that

$$\begin{aligned} F_{xx} &= O(1)(F_t + (a-a_1)\tilde{w}_t + F\tilde{w}_t + F_x^2 \\ &\quad + FF_x\tilde{w}_x + (\tilde{w}_{xx} + \tilde{w}_x^2)F^2). \end{aligned} \tag{2.72}$$

Taking the L^2-norm in (2.72), we have

$$\begin{aligned} w_3(t)\|F_{xx}\|^2 &\leq Cw_3(t)(\|F_t\|^2 + \|(a-a_1)\tilde{w}_t\|^2 + \|F\tilde{w}_t\|^2 + \|F_x^2\|^2 \\ &\quad + \|FF_x\tilde{w}_x\|^2 + \|(\tilde{w}_{xx} + \tilde{w}_x^2)F^2\|^2) \\ &\leq C(1 + w_3(t)\|F_x^2\|^2) \\ &\leq C(1 + w_3(t)\|F_x\|^2(\|F_x\|^2 + \|F_{xx}\|^2)) \\ &\leq C + Cw_3(t)\|F_x\|^2\|F_{xx}\|^2 \end{aligned}$$

which implies

$$w_3(t)\|F_{xx}\|^2 \leq C.$$

Then

$$\|F_x\|^2_{L^\infty} \leq Cw_3(t)^{-\frac{1}{2}}w_2(t)^{-\frac{1}{2}}.$$

Last, if we take the L^∞-norm in (2.72), we obtain

$$\begin{aligned} \|F_{xx}\|_{L^\infty} &\leq C(\|F_t\|_{L^\infty} + \|(a-a_1)\tilde{w}_t\|_{L^\infty} + \|F_x\|^2_{L^\infty} \\ &\quad + \|FF_x\tilde{w}_x\|_{L^\infty} + \|F^2(\tilde{w}_{xx} + \tilde{w}_x^2)\|_{L^\infty}) \\ &\leq C(w_3(t)w_4(t))^{-\frac{1}{4}}. \end{aligned}$$

□

Then Theorem 2.2 follows from Theorem 2.8, Lemmas 2.9–2.10, and Corollary 2.11.

Now, from F it is easy to obtain the solution ϕ of (2.4) and from ϕ the unique smooth solution w of (2.2). By defining $\tilde{v} = a^{-1}(x)w$ and $\tilde{u} = -(w^{-\gamma})_x$, we obtain the solution of (1.4). Theorem 1.1 then follows from Theorem 2.2 and the decay estimates follow from the interpolation inequality and (2.35).

2.2. Case 2: $\boldsymbol{w_- = w_+ = \mathrm{const} = \bar{w}}$. Observe that $w = \bar{w}$ is a stationary solution to (2.2). We will prove Theorem 1.2 by solving the Cauchy problem (2.2) near $\bar{w}$.

Let us denote by $\tilde{\phi} = w - \bar{w}$, then $\tilde{\phi}$

$$\begin{cases} \tilde{\phi}_t - ba(x)\tilde{\phi}_{xx} + a(x)(f_1(\tilde{\phi})\tilde{\phi}^2)_{xx} = 0, \\ \tilde{\phi}(x,0) = \tilde{\phi}_0(x) = w_0(x) - \bar{w}, \end{cases} \tag{2.73}$$

where

$$b = \gamma \bar{w}^{-(\gamma+1)}, \ f_1\tilde{\phi}^2 = (\bar{w} + \tilde{\phi})^{-\gamma} - \bar{w}^{-\gamma} - b\tilde{\phi}.$$

Then we have the following.

Theorem 2.12. *Suppose $\tilde{\phi}_0(x)$ and $s(x) = s_0(x)$ are C^2 functions and $\tilde{\phi}_0 \in H^2(R)$. There exists $\varepsilon_0 > 0$ such that if $0 < \varepsilon < \varepsilon_0$ and $\|\tilde{\phi}_0\|_{H^2} \le \varepsilon$, then* (2.73) *has a unique global smooth solution $\tilde{\phi}(x,t)$ satisfying*

$$\|\tilde{\phi}(\cdot,t)\|_{H^2}^2 + \int_0^t \|\tilde{\phi}_x(\cdot,\tau)\|_{H^2}^2 \, d\tau \le C\varepsilon^2 \tag{2.74}$$

and

$$\lim_{t\to\infty} \|\tilde{\phi}(\cdot,t)\|_{L^\infty} = 0.$$

Furthermore, if $\tilde{\phi}_0 \in L^1(R)$, then

$$\sum_{j=0}^{2} w_{j+1}(t)\|\partial_x^j \tilde{\phi}(\cdot,t)\|^2 + \int_0^t \sum_{j=1}^{3} w_j(\tau)\|\partial_x^j \tilde{\phi}(\cdot,\tau)\|^2 \, d\tau \le C.$$

Proof. Since the local result for (2.73) is classical, to prove the first part of Theorem 2.12, it is sufficient to derive the uniform estimate (2.74) under the a priori assumption $\|\tilde{\phi}\|_{H^2} \le \delta_0$ for δ_0 suitably small.

Multiply $(2.73)_1$ by $a^{-1}\tilde{\phi}$, integrate it over $(-\infty,+\infty) \times [0,t]$, and one then has

$$\begin{aligned} &\|\tilde{\phi}(\cdot,t)\|^2 + \int_0^t \|\tilde{\phi}_x(\cdot,\tau)\|^2 \, d\tau \\ &\le C\varepsilon_0^2 + C\left|\int_0^t \int_{-\infty}^{+\infty} (f_1(\tilde{\phi})\tilde{\phi}^2)_x \tilde{\phi}_x \, dxd\tau\right| \\ &\le C\varepsilon_0^2 + C\delta_0 \int_0^t \|\tilde{\phi}_x\|^2 \, d\tau, \end{aligned}$$

which implies

$$\|\tilde{\phi}(\cdot,t)\|^2 + \int_0^t \|\tilde{\phi}_x(\cdot,\tau)\|^2 \, d\tau \le C\varepsilon_0^2. \tag{2.75}$$

Next, multiplying $(2.73)_1$ by $\tilde{\phi}_{xx}$ and integrating over $(-\infty,+\infty)\times[0,t]$, one has

$$\begin{aligned}&\|\tilde{\phi}_x(\cdot,t)\|^2+\int_0^t\|\tilde{\phi}_{xx}(\cdot,\tau)\|^2\ d\tau\\&\leq C\varepsilon_0^2+C\left|\int_0^t\int_{-\infty}^{+\infty}(f_1(\tilde{\phi})\tilde{\phi}^2)_{xx}\tilde{\phi}_{xx}\ dxd\tau\right|\\&\leq C\varepsilon_0^2+\frac{1}{2}\int_0^t\|\tilde{\phi}_{xx}(\cdot,\tau)\|^2\ d\tau+C\int_0^t\int_{-\infty}^{+\infty}(f_1(\tilde{\phi})\tilde{\phi}^2)_{xx}^2\ dxd\tau\\&\leq C\varepsilon_0^2+\left(\frac{1}{2}+C\delta_0\right)\int_0^t\|\tilde{\phi}_{xx}(\cdot,\tau)\|^2\ d\tau.\end{aligned}$$

Thus we get

$$\|\tilde{\phi}_x(\cdot,t)\|^2+\int_0^t\|\tilde{\phi}_{xx}(\cdot,\tau)\|^2\ d\tau\leq C\varepsilon_0^2. \tag{2.76}$$

Finally, by differentiating $(2.73)_1$ in x and by repeating the previous procedure, one can derive

$$\|\tilde{\phi}_{xx}(\cdot,t)\|^2+\int_0^t\|\tilde{\phi}_{xxx}(\cdot,\tau)\|^2\ d\tau\leq C\varepsilon_0^2. \tag{2.77}$$

The estimate (2.74) follows from (2.75)–(2.77) and then (2.73) has a unique smooth solution $\tilde{\phi}$ such that

$$\lim_{t\to\infty}\|\tilde{\phi}(\cdot,t)\|_{L^\infty}=0.$$

We proceed now to prove the second part of Theorem 2.12. In this framework, we can develop a theory similar to what we did in the previous sections. Actually, it is less complicated since $\bar{w}$ is a constant.

Observe that if $\tilde{\phi}_0(x)\in L^1$, we can use the same argument used in Lemma 2.3 to prove

$$\|\tilde{\phi}(\cdot,t)\|_{L^1}\leq\|\tilde{\phi}_0\|_{L^1}. \tag{2.78}$$

Then we can employ the same argument used in subsection 2.1 to complete the proof of the decay estimates. We perform here just the first two orders estimates.

For the first order estimates, we multiply $(2.73)_1$ by $w_1(t)a^{-1}\tilde{\phi}$ then integrate it by parts over $(-\infty,+\infty)\times[0,t]$. We have

$$\begin{aligned}&w_1(t)\|\tilde{\phi}(\cdot,t)\|^2+\int_0^t w_1(\tau)\|\tilde{\phi}_x(\cdot,\tau)\|^2\ d\tau\\&\leq C+C\int_0^t(1+\tau)^{-1}w_1(\tau)\|\tilde{\phi}(\cdot,\tau)\|_{L^\infty}\|\tilde{\phi}(\cdot,\tau)\|_{L^1}\ d\tau\\&\quad+C\delta_0\int_0^t w_1(\tau)\|\tilde{\phi}_x(\cdot,\tau)\|^2\ d\tau,\end{aligned}\tag{2.79}$$

where

$$\begin{aligned}&\int_0^t(1+\tau)^{-1}w_1(\tau)\|\tilde{\phi}(\cdot,\tau)\|_{L^\infty}\|\tilde{\phi}(\cdot,\tau)\|_{L^1}\ d\tau\\&\leq C(\varepsilon_1)\int_0^t w_1(\tau)(1+t)^{-\frac{3}{2}}\ d\tau+\varepsilon_1\int_0^t w_1(\tau)\|\tilde{\phi}_x(\cdot,\tau)\|^2\ d\tau.\end{aligned}\tag{2.80}$$

By choosing ε_1 small, we see from (2.79)–(2.80)

$$w_1(t)\|\tilde{\phi}(\cdot,t)\|^2 + \int_0^t w_1(\tau)\|\tilde{\phi}_x(\cdot,\tau)\|^2\,d\tau \le C. \tag{2.81}$$

For the second order estimates, we multiply $(2.73)_1$ by $w_2(t)\tilde{\phi}_{xx}$, integrate it by parts over $(-\infty,+\infty)\times[0,t]$, then

$$\begin{aligned}
&w_2(t)\|\tilde{\phi}_x(\cdot,t)\|^2 + \int_0^t w_2(\tau)\|\tilde{\phi}_{xx}(\cdot,\tau)\|^2\,d\tau\\
&\le C + \frac{1}{2}\int_0^t w_2(\tau)\|\tilde{\phi}_{xx}(\cdot,\tau)\|^2\,d\tau\\
&\quad + C\int_0^t\int_{-\infty}^{+\infty} w_2(\tau)(f_1(\tilde{\phi})\tilde{\phi}^2)_{xx}^2\,dxd\tau,
\end{aligned} \tag{2.82}$$

where

$$\begin{aligned}
&\int_0^t\int_{-\infty}^{+\infty} w_2(\tau)(f_1(\tilde{\phi})\tilde{\phi}^2)_{xx}^2\,dxd\tau\\
&\le C\delta_0\int_0^t w_2(\tau)\|\tilde{\phi}_{xx}(\cdot,\tau)\|^2\,d\tau\\
&\quad + C\int_0^t w_2(\tau)\|\tilde{\phi}_x(\tau)\|^2(\|\tilde{\phi}_x\|^2+\|\tilde{\phi}_{xx}\|^2)\,d\tau\\
&\le C\delta_0\int_0^t w_2(\tau)\|\tilde{\phi}_{xx}(\cdot,\tau)\|^2\,d\tau + C\int_0^t w_2(\tau)\|\tilde{\phi}_x(\tau)\|^4\,d\tau.
\end{aligned} \tag{2.83}$$

We conclude from (2.82)–(2.83) that

$$\begin{aligned}
&w_2(t)\|\tilde{\phi}_x(\cdot,t)\|^2 + \int_0^t w_2(\tau)\|\tilde{\phi}_{xx}(\cdot,\tau)\|^2\,d\tau\\
&\le C + C\int_0^t w_2(\tau)\|\tilde{\phi}_x(\tau)\|^4\,d\tau,
\end{aligned}$$

which, together with the help of Gronwall inequality, implies that

$$w_2(t)\|\tilde{\phi}_x(\cdot,t)\|^2 + \int_0^t w_2(\tau)\|\tilde{\phi}_{xx}(\cdot,\tau)\|^2\,d\tau \le C. \qquad \square \tag{2.84}$$

In the following, we denote by $v_2(x,t)=a^{-1}w(x,t)$ the solution to (2.1) obtained in Theorem 1.2, and $u_2(x,t)=-p(v_2,s)_x$. Theorem 1.2 then follows from Theorem 2.12.

3. Convergence to similarity solutions. In this section, we will study (1.2)–(1.3) for Case 1, namely, we assume that $s_-=s_+=\bar{s}$. We shall prove Theorem 1.3 by comparing the solutions of (1.2)–(1.3) with those of (1.4) obtained in Theorem 2.2. Since the result for $s(x,t)$ is clear, in the following part we only deal with $(v,u)(x,t)$.

Let $(\tilde{v},\tilde{u},s(x))$ be the solution of (1.4) with the initial data $(\tilde{v}_0(x),s_0(x))$. As pointed in introduction, we will only prove Theorem 1.3 for the case where $u_-=$

$u_+ = 0$ and thus (1.14) turns into

$$\int_{-\infty}^{+\infty} (v_0(x) - \tilde{v}_0(x))\ dx = 0. \tag{3.1}$$

Let us denote

$$v_e = v - \tilde{v},\ u_e = u - \tilde{u}, \tag{3.2}$$

then it follows from (1.2) and (1.4) that

$$\begin{cases} v_{et} - u_{ex} = 0, \\ u_{et} + [p(\tilde{v} + v_e, s) - p(\tilde{v}, s)]_x = -u_e + p(\tilde{v}, s)_{xt}. \end{cases} \tag{3.3}$$

As usual let us consider

$$y = \int_{-\infty}^{x} v_e(\xi)\ d\xi, \tag{3.4}$$

which satisfies the following nonlinear wave equation:

$$\begin{cases} y_{tt} + [p(y_x + \tilde{v}, s) - p(\tilde{v}, s)]_x + y_t = p(\tilde{v}, s)_{xt}, \\ y(x, 0) = y_0(x) = \displaystyle\int_{-\infty}^{x} (v_0 - \tilde{v}_0)(\xi)\ d\xi, \\ y_t(x, 0) = y_1(x) = u_0(x) - \tilde{u}(x, 0) \end{cases}$$

since $y_x = v_e$ and $y_t = u_e$. Therefore

$$\begin{cases} y_{tt} + (p_v(\tilde{v}, s)y_x)_x + y_t = p(\tilde{v}, s)_{xt} - (F_1(\tilde{v}, y_x, s)y_x^2)_x, \\ y(x, 0) = y_0(x) = \displaystyle\int_{-\infty}^{x} (v_0 - \tilde{v}_0)(\xi)\ d\xi, \\ y_t(x, 0) = y_1(x) = u_0(x) - \tilde{u}(x, 0), \end{cases} \tag{3.5}$$

where

$$p(y_x + \tilde{v}, s) - p(\tilde{v}, s) = p_v(\tilde{v}, s)y_x + F_1(\tilde{v}, y_x, s)y_x^2.$$

The main result of this section is the following.

THEOREM 3.1. *There exists $\delta_0 > 0$ such that if $0 < \delta < \delta_0$ and*

$$\|y_0\|_{H^3} + \|y_1\|_{H^2} + |v_+ - v_-| \le \delta,$$

then (3.5) has a unique smooth solution $y \in H^3$ and $y_t \in H^2$ satisfying

$$\|y(t)\|_{H^3}^2 + \|y_t(t)\|_{H^2}^2 + \int_0^t \|(y_x, y_t)(\tau)\|_{H^2}\ d\tau \le C\delta^2.$$

Moreover,

$$(1 + t)\|y_x(\cdot, t)\|^2 + (1 + t)^2\|y_t(\cdot, t)\|^2 \le C \tag{3.6}$$

and

$$\|y_x(\cdot,t)\|_{L^\infty} \le C(1+t)^{-\frac{3}{4}}, \ \|y_t(\cdot,t)\|_{L^\infty} \le C(1+t)^{-\frac{5}{4}}. \tag{3.7}$$

Then Theorem 1.3 follows from Theorem 2.2 and Theorem 3.1.

We now prove Theorem 3.1. First of all, we have the following.

THEOREM 3.2. *There exists* $\delta_0 > 0$ *such that if* $0 < \delta < \delta_0$ *and*

$$\|y_0\|_{H^3} + \|y_1\|_{H^2} + |v_+ - v_-| \le \delta,$$

then (3.5) *has a unique global smooth solution* $y \in H^3$ *and* $y_t \in H^2$ *satisfying*

$$\|y(t)\|_{H^3}^2 + \|y_t(t)\|_{H^2}^2 + \int_0^t \|(y_x, y_t)(\tau)\|_{H^2}\ d\tau \le C\delta^2. \tag{3.8}$$

Proof. It is sufficient to prove the uniform estimates (3.8) under the following a priori assumption:

$$\|y(t)\|_{H^3}^2 + \|y_t\|_{H^2}^2 \le \varepsilon$$

for $\varepsilon > 0$ suitably small.

Multiplying $(3.5)_1$ by $y + 2y_t$, we have

$$\begin{aligned}&\left[y_t^2 - p_v(\tilde{v}, s)y_x^2 + \frac{1}{2}y^2 + yy_t\right]_t + y_t^2 - p_v(\tilde{v}, s)y_x^2 \\ &= p_{vv}(\tilde{v}, s)\tilde{v}_t(y_x^2 - y_x - 2y_{xt}) + (F_1 y_x^2 - p(\tilde{v}, s)_t)(y_x + 2y_{tx}) + \{\cdots\}_x,\end{aligned} \tag{3.9}$$

where $\{\cdots\}_x$ denote the terms which disappear after integration with respect to x. Integrating (3.9) over $[0,t] \times (-\infty, +\infty)$, we get

$$\begin{aligned}&\|(y, y_t, y_x)(t)\|^2 + \int_0^t \|(y_t, y_x)(\tau)\|^2\ d\tau \\ &\le C\delta^2 + C\left|\int_0^t \int_{-\infty}^{+\infty} p_{vv}(\tilde{v}, s)\tilde{v}_t y_x^2 \right. \\ &\qquad \left. -p(\tilde{v}, s)_t(y_x + 2y_{xt}) + F_1 y_x^2(y_x + 2y_{tx})\ dxd\tau\right|.\end{aligned} \tag{3.10}$$

Due to the smallness of ε, we can reduce (3.10) into

$$\begin{aligned}&\|(y, y_t, y_x)(t)\|^2 + \int_0^t \|(y_t, y_x)(\tau)\|^2\ d\tau \\ &\le C\delta^2 + C\left|\int_0^t \int_{-\infty}^{+\infty} p(\tilde{v}, s)_t(y_x + 2y_{xt})\ dxd\tau\right|,\end{aligned} \tag{3.11}$$

while

$$\begin{aligned}&\left|\int_0^t \int_{-\infty}^{+\infty} p(\tilde{v}, s)_t(y_x + 2y_{xt})\ dxd\tau\right| \\ &\le C(\varepsilon_1)\int_0^t \int_{-\infty}^{+\infty} (p(\tilde{v}, s)_t^2 + p(\tilde{v}, s)_{tx}^2)\ dxd\tau + \varepsilon_1 \int_0^t \int_{-\infty}^{+\infty} (y_x^2 + y_t^2)\ dxd\tau\end{aligned} \tag{3.12}$$

and

$$\int_0^t \int_{-\infty}^{+\infty} p(\tilde{v},s)_t^2 + p(\tilde{v},s)_{tx}^2 \, dxd\tau \le C\delta^2. \tag{3.13}$$

Now, by taking ε_1 small, it reads from (3.11)–(3.13) that

$$\|(y,y_t,y_x)(t)\|^2 + \int_0^t \|(y_x,y_t)(\tau)\|^2 \, d\tau \le C\delta^2. \tag{3.14}$$

We now differentiate (3.5) in x and then

$$y_{ttx} + (p_v(\tilde{v},s)y_x)_{xx} + y_{tx} = p(\tilde{v},s)_{xtx} - (F_1(\tilde{v},y_x,s)y_x^2)_{xx}. \tag{3.15}$$

If we multiply (3.15) by y_x+2y_{tx} and integrate the resulting equation over $[0,1]\times[0,t]$, by using (3.14) we get

$$\begin{aligned}
&\|(y_x,y_{tx},y_{xx})(t)\|^2 + \int_0^t \|(y_{tx},y_{xx})(\tau)\|^2 \, d\tau \\
&\le C\delta^2 + C\int_0^t \int_{-\infty}^{+\infty} (F_1y_x^2)_x^2 \, dxd\tau \\
&\qquad + C\int_0^t \int_{-\infty}^{+\infty} [O(1)y_xy_{xx}^2]_t \, dxd\tau \\
&\qquad + C\delta\int_0^t \int_{-\infty}^{+\infty} (y_x^2+y_{tx}^2+y_{xx}^2) \, dxd\tau,
\end{aligned}$$

which implies

$$\|(y_x,y_{tx},y_{xx})(t)\|^2 + \int_0^t \|(y_{tx},y_{xx})(\tau)\|^2 \, d\tau \le C\delta^2.$$

Repeating the above procedure, we can easily obtain the third order estimates and complete the proof of this theorem. □

With the help of Theorem 3.2, it is easy to obtain the following convergence results by using an argument similar to the proof of Theorem 2.8.

THEOREM 3.3. *The solution y to* (3.5) *in the Theorem* 3.2 *satisfies*

$$\lim_{t\to\infty} (\|y(\cdot,t)\|_{L^\infty} + \|(y_t,y_x)(\cdot,t)\|_{H^1}) = 0.$$

We investigate now the problem of the decay rate. We will follow the approach introduced by [19] concerning the isentropic case. However, since the entropy $s(x)$ is not constant here, some modifications are necessary.

LEMMA 3.4. *Under the previous hypotheses, it follows that*

$$(1+t)\|(y_x,y_t)(t)\|^2 + \int_0^t (1+\tau)\|y_t(\tau)\|^2 \, d\tau \le C\delta^2.$$

Proof. First, we notice that $(3.5)_1$ is equivalent to

$$y_{tt} + y_t + [p(\tilde{v}+y_x,s) - p(\tilde{v},s)]_x = p(\tilde{v},s)_{xt}. \tag{3.16}$$

Multiplying (3.16) by $(1+t)y_t$, after some calculations we get

$$\begin{aligned}&\left[(1+t)\left(\frac{1}{2}y_t^2+q\right)\right]_t+(1+t)y_t^2-q\\ &-\int_0^{y_x}\left[p_v(\tilde{v}+\xi,s)-p_v(\tilde{v},s)\right]\,d\xi+\tilde{v}_t(1+t)y_x^2-\frac{1}{2}y_t^2\\ &=(1+t)y_tp(\tilde{v},s)_{xt}+\{\cdots\}_x.\end{aligned}\tag{3.17}$$

Integrating (3.17) over $[0,t]\times(-\infty,+\infty)$, with the help of (3.8), we have

$$\begin{aligned}&(1+t)\|(y_x,y_t)(t)\|^2+\int_0^t(1+\tau)\|y_t(\tau)\|^2\,d\tau\\ &\le C\delta^2+\frac{1}{2}\int_0^t(1+\tau)\|y_t(\tau)\|^2\,d\tau,\end{aligned}$$

which implies

$$(1+t)\|(y_t,y_x)(t)\|^2+\int_0^t(1+\tau)\|y_t(\tau)\|^2\,d\tau\le C\delta^2.\tag{3.18}$$

Here we have used the following properties:

$$\begin{aligned}&q=-\int_0^{y_x}\left[p(\tilde{v}+\xi,s)-p(\tilde{v},s)\right]\,d\xi=O(1)y_x^2,\\ &\int_0^{y_x}\left[p_v(\tilde{v}+\xi,s)-p_v(\tilde{v},s)\right]\,d\xi=O(1)y_x^2,\\ &\tilde{v}_t\le O(1)(F_t+\tilde{w}_t)\le O(1)(1+t)^{-1}.\qquad\square\end{aligned}$$

LEMMA 3.5. *Under the previous hypotheses, we have*

$$(1+t)^2\|(y_t,y_{tt},y_{tx})(t)\|^2+\int_0^t(1+\tau)^2\|(y_{tt},y_{tx})(\tau)\|^2\,d\tau\le C\delta^2.$$

Proof. Differentiating $(3.5)_1$ in t, we have

$$y_{ttt}+(p_v(\tilde{v},s)y_x)_{xt}+y_{tt}=p(\tilde{v},s)_{xtt}-(F_1y_x^2)_{xt}.\tag{3.19}$$

Let us multiply (3.19) by $(1+t)y_t$ and $(1+t)y_{tt}$, respectively, then we deduce

$$\begin{aligned}&\left[(1+t)\left(y_ty_{tt}+\frac{1}{2}y_t^2\right)\right]_t-p_v(\tilde{v},s)(1+t)y_{tx}^2-(1+t)y_{tt}^2-\frac{1}{2}y_t^2-y_ty_{tt}\\ &=p_{vv}\tilde{v}_t(1+t)y_xy_{tx}+(1+t)y_t(p(\tilde{v},s)_{xtt}-(F_1y_x^2)_{xt})+\{\cdots\}_x,\end{aligned}\tag{3.20}$$

$$\begin{aligned}&\left[\frac{1}{2}(1+t)(y_{tt}^2-p_vy_{tx}^2)\right]_t+(1+t)y_{tt}^2\\ &-\frac{1}{2}y_{tt}^2+\frac{1}{2}p_vy_{tx}^2+\frac{1}{2}(1+t)p_{vv}\tilde{v}_ty_{tx}^2+(1+t)y_{tt}(y_xp_{vv}\tilde{v}_t)_x\\ &=(1+t)y_{tt}(p(\tilde{v},s)_{xtt}-(F_1y_x^2)_{xt})+\{\cdots\}_x.\end{aligned}\tag{3.21}$$

By using Theorem 3.2 and Lemma 3.4 and by integrating $8 \times (3.21) + (3.20)$ one has

$$\begin{aligned}
&(1+t)\|(y_t, y_{tt}, y_{tx})\|^2 + \int_0^t (1+\tau)\|(y_{tt}, y_{tx})(\tau)\|^2 \, d\tau \\
&\le C\left(\delta^2 + \int_0^t \int_{-\infty}^{+\infty} (1+\tau) p(\tilde{v}, s)_{xtt}^2 \, dx d\tau \right. \\
&\qquad \left. + \left| \int_0^t \int_{-\infty}^{+\infty} (1+\tau)(y_{tx} + y_{ttx})(F_1 y_x^2)_t \, dx d\tau \right| \right) \\
&\le C\delta^2 + C\left| \int_0^t \int_{-\infty}^{+\infty} (1+\tau)(y_{tx} + y_{ttx})(F_1 y_x^2)_t \, dx d\tau \right|.
\end{aligned} \tag{3.22}$$

Moreover, one has

$$\begin{aligned}
&\left| \int_0^t \int_{-\infty}^{+\infty} (1+\tau) y_{tx} (F_1 y_x^2)_t \, dx d\tau \right| \\
&\le C \int_0^t \int_{-\infty}^{+\infty} (1+\tau)(|y_x| y_{tx}^2 + |\tilde{v}_t y_x^2 y_{tx}|) \, dx d\tau \\
&\le C\delta^2 + C\delta \int_0^t \int_{-\infty}^{+\infty} (1+\tau) y_{tx}^2 \, dx d\tau
\end{aligned} \tag{3.23}$$

and

$$\begin{aligned}
&\left| \int_0^t \int_{-\infty}^{+\infty} (1+\tau) y_{ttx} (F_1 y_x^2)_t \, dx d\tau \right| \\
&\le C\delta^2 + C\delta \left((1+t)\|y_{tx}(t)\|^2 + \int_0^t (1+\tau)\|y_{tx}\|^2 \, d\tau \right).
\end{aligned} \tag{3.24}$$

In view of the smallness of δ, from (3.22)–(3.24) we have

$$(1+t)\|(y_t, y_{tt}, y_{tx})(t)\|^2 + \int_0^t (1+\tau)\|(y_{tt}, y_{tx})(\tau)\|^2 \, d\tau \le C\delta^2. \tag{3.25}$$

Now we multiply (3.19) by $(1+t)^2 y_t$ and $(1+t)^2 y_{tt}$ and repeat the previous calculations to conclude Lemma 3.5. □

LEMMA 3.6. *The solution y to* (3.5) *in Theorem* 3.2 *satisfies*

$$(1+t)^2\|(V_t, V_x)(t)\|^2 + \int_0^t (1+\tau)\|(V_t, V_x)(\tau)\|^2 \, d\tau \le C\delta^2,$$

where $V = p_v(\tilde{v}, s) y_x$.

Proof. The estimate for V_t can be obtained from Lemma 3.5 and the following relation:

$$V_t = p_v(\tilde{v}, s) y_{tx} + p_{vv}(\tilde{v}, s) \tilde{v}_t y_x.$$

It is easy to see that

$$V_x = -(y_{tt} + y_t + p(\tilde{v}, s)_{xt} + (F_2 V^2)_x), \tag{3.26}$$

where $F_2V^2 = F_1y_x^2$. Then we calculate the decay rate for V_x by using (3.26). First of all, it is easy to see by taking the L^2-norm in (3.26) that

$$(1+t)\|V_x(t)\|^2 \le C\delta^2 \tag{3.27}$$

and

$$\begin{aligned}&(1+t)^2\|V_x\|^2\\ &\le C(1+t)^2(\|y_{tt}\|^2 + \|y_t\|^2 + \|(F_2V^2)_x\|^2)\\ &\le C\delta^2 + \frac{1}{2}(1+t)^2\|V_x\|^2 + C(1+t)(\|V\|^2 + \|V_x\|^2)\\ &\le C\delta^2 + \frac{1}{2}(1+t)^2\|V_x\|^2\end{aligned}$$

thus

$$(1+t)^2\|V_x\|^2 \le C\delta^2. \tag{3.28}$$

Then, multiplying (3.26) by $(1+t)V_x$ and integrating it, one has

$$\begin{aligned}&\int_0^t (1+\tau)\|V_x(\tau)\|^2\ d\tau\\ &\le C\int_0^t\int_{-\infty}^{+\infty}(1+\tau)(y_{tt}^2 + y_t^2 + p(\tilde{v},s)_{xt}^2 + (F_2V^2)_x^2)\ dxd\tau\\ &\le C\delta^2 + C\int_0^t\int_{-\infty}^{+\infty}(F_2V^2)_x^2(1+\tau)\ dxd\tau\\ &\le C\delta^2 + C\int_0^t\int_{-\infty}^{+\infty}(1+\tau)V^4\ dxd\tau\\ &\le C\delta^2 + C\int_0^t(\|y_x\|^2 + \|y_{xx}\|^2)(1+\tau)\|y_x\|^2\ d\tau\\ &\le C\delta^2. \quad \square\end{aligned} \tag{3.29}$$

The following result easily holds by repeating the previous arguments on the equation (3.19) differentiated with respect to x.

Lemma 3.7. *The solution y to* (3.5) *in Theorem* 3.2 *satisfies*

$$(1+t)^2\|(y_{ttx}, y_{txx})(t)\|^2 + \int_0^t (1+\tau)^2\|(y_{ttx}, y_{txx})(\tau)\|^2\ d\tau \le C\delta^2.$$

Now we can prove the desired estimates on y_{tx}.

Lemma 3.8. *Under the previous hypotheses, one has*

$$(1+t)^3\|(y_{tt}, y_{tx})(t)\|^2 + \int_0^t (1+\tau)^3\|y_{tt}(\tau)\|^2\ d\tau \le C\delta^2.$$

Proof. Multiply (3.19) by $(1+t)^3y_{tt}$, then we obtain

$$\begin{aligned}&(1+t)^3\|(y_{tt}, y_{tx})(t)\|^2 + \int_0^t (1+\tau)^3\|y_{tt}(\tau)\|^2\ d\tau\\ &\le C\delta^2 + C\left|\int_0^t\int_{-\infty}^{+\infty}(1+\tau)^3y_{ttx}(F_1y_x^2)_t\ dxd\tau\right|.\end{aligned} \tag{3.30}$$

We have that

$$\tag{3.31}\begin{aligned}&\left|\int_0^t\int_{-\infty}^{+\infty}(1+\tau)^3 y_{ttx}(F_1y_x^2)_t\ dxd\tau\right|\\&\le C\left|\int_0^t\int_{-\infty}^{+\infty}[O(1)(1+\tau)^3y_{tx}^2y_x]_t\ dxd\tau\right|\\&+C\left|\int_0^t\int_{-\infty}^{+\infty}((1+t)^2y_{tx}^2+(1+\tau)^3y_{tx}^3\ dxd\tau\right|\\&+C\left|\int_0^t\int_{-\infty}^{+\infty}(1+\tau)^3y_{tt}(F_3\tilde{v}_tV^2)_x\ dxd\tau\right|\\&\le C(\alpha_1)\delta^2+C\delta(1+t)^3\|y_{tx}(t)\|^2+\alpha_1\int_0^t(1+\tau)^3\|y_{tt}\|^2\ d\tau,\end{aligned}$$

where $F_3V^2 = F_{1v}y_x^2$.

By choosing α_1 suitable small, we conclude from (3.30)–(3.31) that

$$(1+t)^3\|(y_{tt},y_{tx})(t)\|^2+\int_0^t(1+\tau)^3\|y_{tt}\|^2\ d\tau\le C\delta^2. \qquad \square$$

Therefore, we obtain the following desired decay rates.

THEOREM 3.9. *The solution y to* (3.5) *in Theorem* 3.2 *satisfies*

$$\tag{3.32}\sum_{k=0}^{1}[(1+t)^{k+1}\|\partial_x^kV(\cdot,t)\|^2+(1+t)^{k+2}\|\partial_x^ky_t(\cdot,t)\|^2]\le C$$

and

$$\tag{3.33}\|y_x(\cdot,t)\|_{L^\infty}\le C(1+t)^{-\frac{3}{4}},\ \ \|y_t(\cdot,t)\|_{L^\infty}\le C(1+t)^{-\frac{5}{4}}.$$

Proof. (3.32) comes directly from Lemmas 3.4–3.8. (3.33) follows from the interpolation inequality and (3.32), where

$$\|y_x(\cdot,t)\|_{L^\infty}\le C\|V(\cdot,t)\|_{L^\infty}\le C(1+t)^{-\frac{3}{4}}. \qquad \square$$

Theorem 3.1 then follows from Theorem 3.2 and Theorem 3.9.

4. Convergence to stationary solution. This section is devoted to proving Theorem 4 and Theorem 5, where (v_-,v_+) and (s_-,s_+) are chosen so that $p(v_-,s_-) = p(v_+,s_+) = \bar{p} = \text{const}$.

Denote by $(v_2,u_2)(x,t)$ the solution of (1.4) obtained in Theorem 2.12. We solve (1.2)–(1.3) near $(v_2,u_2)(x,t)$ under $u_- = u_+ = 0$ and then (1.16) implies

$$\tag{4.1}\int_{-\infty}^{+\infty}(v_0(x)-v_2(x,0))\ dx=0.$$

Similarly to section 3, we set

$$\tilde{y}=\int_{-\infty}^{x}(v(\xi,t)-v_2(\xi,t))\ d\xi,$$

which satisfies

$$\begin{cases} \tilde{y}_{tt} + (p_v(v_2,s)\tilde{y}_x)_x + \tilde{y}_t = p(v_2,s)_{xt} - (F_1(v_2,\tilde{y}_x,s)\tilde{y}_x^2)_x, \\ \tilde{y}(x,0) = \tilde{y}_0(x) = \displaystyle\int_{-\infty}^{x} (v_0(\xi) - v_2(\xi,0))\, d\xi, \\ \tilde{y}_t(x,0) = \tilde{y}_1(x) = u_0(x) - u_2(x,0), \end{cases} \tag{4.2}$$

where

$$p(\tilde{y}_x + v_2, s) - p(v_2,s) = p_v(v_2,s)\tilde{y}_x + F_1(v_2,\tilde{y}_x,s)\tilde{y}_x^2.$$

From the results in subsection 2.2 and the argument used in section 3, it is clear that the same argument of the section 3 can be used here to prove the following results.

THEOREM 4.1. *There exists* $\delta_0 > 0$ *such that if* $0 < \delta < \delta_0$ *and*

$$\|\tilde{y}_0\|_3^2 + \|\tilde{y}_1\|_2^2 \le \delta,$$

then (4.2) *has a unique smooth solution* $\tilde{y}$ *satisfying*

$$\sum_{k=0}^{1}[(1+t)^{k+1}\|\partial_x^k V_1(\cdot,t)\|^2 + (1+t)^{k+2}\|\partial_x^k \tilde{y}_t(\cdot,t)\|^2] \le C, \tag{4.3}$$

with $V_1 = p_v(v_2,s)\tilde{y}_x$, *and*

$$\|\tilde{y}_x(\cdot,t)\|_{L^\infty} \le C(1+t)^{-\frac{3}{4}},\ \|\tilde{y}_t(\cdot,t)\|_{L^\infty} \le C(1+t)^{-\frac{5}{4}}. \tag{4.4}$$

It is clear that Theorem 1.4 comes from Theorem 1.2 and Theorem 4.1.

We turn to proving Theorem 1.5 next. Since $(v_1(x), 0, s_0(x))$ is the stationary solution of both (1.2) and (1.4), we can also solve (1.2)–(1.3) near v_1 instead of v_2, under the condition $u_- = u_+ = 0$ and (1.19), then

$$\int_{-\infty}^{+\infty} (v_0(x) - v_1(x))\, dx = 0. \tag{4.5}$$

Denote

$$z = \int_{-\infty}^{x} (v(\xi,t) - v_1(\xi))\, d\xi; \tag{4.6}$$

then it follows that

$$\begin{cases} z_{tt} + (p_v(v_1,s)z_x)_x + z_t = -(F_1(v_1,z_x,s)z_x^2)_x, \\ z(x,0) = z_0(x) = \displaystyle\int_{-\infty}^{x} (v_0(\xi) - v_1(\xi))\, d\xi, \\ z_t(x,0) = z_1(x) = u_0(x), \end{cases} \tag{4.7}$$

where

$$p(z_x + v_1, s) - p(v_1,s) = p_v(v_1,s)z_x + F_1(v_1,z_x,s)z_x^2.$$

We will prove the following theorem.

THEOREM 4.2. *There exists $\delta_0 > 0$ such that if $0 < \delta < \delta_0$ and*

$$\|z_0\|_{H^3} + \|z_1\|_{H^2} \leq \delta,$$

then (4.7) *has a unique global smooth solution z satisfying*

$$\sum_{k=0}^{1}[(1+t)^{k+1}\|\partial_x^k V_2(\cdot,t)\|^2 + (1+t)^{k+2}\|\partial_x^k z_t(\cdot,t)\|^2] \leq C,$$

where $V_2 = p_v(v_1,s)z_x$ and

$$\|z_x(\cdot,t)\|_{L^\infty} \leq C(1+t)^{-\frac{3}{4}},\ \|z_t(\cdot,t)\|_{L^\infty} \leq C(1+t)^{-\frac{5}{4}}.$$

Hence (1.2)–(1.3) *has a unique global smooth solution $(v,u,s)(x,t)$ such that*

$$\|(v-v_1)(\cdot,t)\|_{L^\infty} \leq C(1+t)^{-\frac{3}{4}},\ \|u(\cdot,t)\|_{L^\infty} \leq C(1+t)^{-\frac{5}{4}}.$$

We note that Theorem 4.2 implies Theorem 1.5.

Using the same proof as in Theorem 3.2, noting $p(v_1,s) = \text{const}$, we can deduce the following lemma.

LEMMA 4.3. *There exists $\delta_0 > 0$ such that if $0 < \delta < \delta_0$ and*

$$\|z_0\|_{H^3} + \|z_1\|_{H^2} \leq \delta,$$

then (4.7) *has a unique smooth solution z satisfying*

$$\|z(t)\|_{H^3}^2 + \|z_t(t)\|_{H^2}^2 + \int_0^t \|(z_x,z_t)(\tau)\|_{H^2}\ d\tau \leq C\delta^2. \tag{4.8}$$

The next result concerns the decay rates.

LEMMA 4.4. *The solution z of* (4.7)*, obtained in Lemma* 4.3*, satisfies*

$$\sum_{k=0}^{1}[(1+t)^{k+1}\|\partial_x^k V_2(\cdot,t)\|^2 + (1+t)^{k+2}\|\partial_x^k z_t(\cdot,t)\|^2] \leq C, \tag{4.9}$$

where $V_2 = p_v(v_1,s)z_x$ and

$$\|z_x(\cdot,t)\|_{L^\infty} \leq C(1+t)^{-\frac{3}{4}},\ \|z_t(\cdot,t)\|_{L^\infty} \leq C(1+t)^{-\frac{5}{4}}. \tag{4.10}$$

Proof. We multiply $(4.7)_1$ by $(1+t)z_t$ and integrate it by parts. Then by using (4.8), we obtain, by a calculation similar to that one in the proof of Lemma 3.4, that

$$(1+t)\|(z_x,z_t)(t)\|^2 + \int_0^t (1+\tau)\|z_t(\tau)\|^2\ d\tau \leq C\delta^2. \tag{4.11}$$

Now let us differentiate $(4.7)_1$ in t, then we have

$$z_{ttt} + (p_v(v_1,s)y_x)_{xt} + y_{tt} = -(F_1 z_x^2)_{xt}. \tag{4.12}$$

Multiplying (4.12) by $(1+t)z_t$ and $(1+t)z_{tt}$, respectively, we deduce

$$\begin{aligned}&\left[(1+t)\left(z_t z_{tt}+\frac{1}{2}z_t^2\right)\right]_t - p_v(v_1,s)(1+t)z_{tx}^2-(1+t)z_{tt}^2\\&=\frac{1}{2}z_t^2+z_t z_{tt}-(F_1 z_x^2)_t(1+t)z_{tx}+\{\cdots\}_x,\end{aligned}\tag{4.13}$$

$$\begin{aligned}&\left[\frac{1}{2}(1+t)(z_{tt}^2-p_v z_{tx}^2)\right]_t+(1+t)z_{tt}^2\\&=\frac{1}{2}z_{tt}^2-\frac{1}{2}p_v z_{tx}^2-(1+t)z_{ttx}(F_1 z_x^2)_t)+\{\cdots\}_x.\end{aligned}\tag{4.14}$$

Then by using (4.8) and (4.11), and by integrating $8\times(4.14)+(4.13)$, we have

$$\begin{aligned}&(1+t)\|(z_{tt},z_{tx})\|^2+\int_0^t(1+\tau)\|(z_{tt},z_{tx})(\tau)\|^2\,d\tau\\&\leq C+C\left|\int_0^t\int_{-\infty}^{+\infty}(1+\tau)(z_{tx}+z_{ttx})(F_1 z_x^2)_t\,dxd\tau\right|.\end{aligned}\tag{4.15}$$

We see that

$$\begin{aligned}&\left|\int_0^t\int_{-\infty}^{+\infty}(1+\tau)z_{tx}(F_1 z_x^2)_t\,dxd\tau\right|\\&\leq C\delta\int_0^t(1+\tau)\|z_{tx}(\tau)\|^2\,d\tau,\end{aligned}\tag{4.16}$$

and

$$\begin{aligned}&\left|\int_0^t\int_{-\infty}^{+\infty}(1+\tau)z_{ttx}(F_1 z_x^2)_t\,dxd\tau\right|\\&\leq C\left|\int_0^t\int_{-\infty}^{+\infty}[O(1)|z_x|(1+\tau)z_{tx}^2]_t\,dxd\tau\right|\\&\quad+C\delta\int_0^t\int_{-\infty}^{+\infty}(1+(1+\tau))z_{tx}^2\,dxd\tau.\end{aligned}\tag{4.17}$$

Due to the smallness of δ, from (4.15)–(4.17) we have

$$(1+t)\|(z_{tt},z_{tx})\|^2+\int_0^t(1+\tau)\|(z_{tt},z_{tx})(\tau)\|^2\,d\tau\leq C.\tag{4.18}$$

Now let us multiply (4.12) by $(1+t)^2 z_t$ and $(1+t)^2 z_{tt}$ and repeat the previous calculations, then

$$(1+t)^2\|(z_t,z_{tt},z_{tx})\|^2+\int_0^t(1+\tau)^2\|(z_{tt},z_{tx})(\tau)\|^2\,d\tau\leq C.\tag{4.19}$$

The same proof as used in Lemma 3.6 yields

$$(1+t)^2\|(V_{2t},V_{2x})\|^2+\int_0^t(1+\tau)\|(V_{2t},V_{2x})(\tau)\|^2\,d\tau\leq C.\tag{4.20}$$

By differentiating (4.12) in x, we get

$$(1+t)^2\|(z_{ttx}, z_{txx})\|^2 + \int_0^t (1+\tau)^2\|(z_{ttx}, z_{txx})(\tau)\|^2 \, d\tau \le C, \tag{4.21}$$

and finally, by multiplying (4.12) by $(1+t)^3 z_{tt}$, it follows that

$$\begin{aligned}
&(1+t)^3\|(z_{tt}, z_{tx})\|^2 + \int_0^t (1+\tau)^3\|z_{tt}(\tau)\|^2 \, d\tau \\
&\le C + C\left|\int_0^t \int_{-\infty}^{+\infty} [O(1)|z_x|(1+\tau)^3 z_{tx}^2]_t \, dxd\tau\right| \\
&\quad + C\delta \int_0^t \int_{-\infty}^{+\infty} (1+\tau)^2 z_{tx}^2 \, dxd\tau + C\left|\int_0^t \int_{-\infty}^{+\infty} (1+\tau)^3 z_{tx}^3 \, dxd\tau\right|,
\end{aligned}$$

which implies

$$\begin{aligned}
&(1+t)^3\|(z_{tt}, z_{tx})\|^2 + \int_0^t (1+\tau)^3\|z_{tt}(\tau)\|^2 \, d\tau \\
&\le C + C\left|\int_0^t \int_{-\infty}^{+\infty} (1+\tau)^3 z_{tx}^3 \, dxd\tau\right|.
\end{aligned} \tag{4.22}$$

We have

$$\begin{aligned}
&\left|\int_0^t \int_{-\infty}^{+\infty} (1+\tau)^3 z_{tx}^3 \, dxd\tau\right| \\
&\le C\int_0^t \int_{-\infty}^{+\infty} (1+\tau)^2 z_{tx}^2 + (1+\tau)^4 z_{tx}^4 \, dxd\tau \\
&\le C + C\int_0^t (1+\tau)^2(\|z_{tx}\|^2 + \|z_{txx}\|^2) \, d\tau \\
&\le C.
\end{aligned} \tag{4.23}$$

Then from (4.22)–(4.23), it follows that

$$(1+t)^3\|(z_{tt}, z_{tx})\|^2 + \int_0^t (1+\tau)^3\|z_{tt}(\tau)\|^2 \, d\tau \le C. \tag{4.24}$$

Hence, (4.9) follows from the combination of (4.11), (4.19)–(4.21), and (4.24). The estimate (4.10) follows from (4.9). □

By combining Lemmas 4.3 and 4.4, we complete the proof of the Theorem 4.2.

Acknowledgments. This work was completed when the second author was visiting the Dipartmento di Matematica Pura ed Applicata, Universita degli Studi di L'Aquila; he wishes to thank Professor Pierangelo Marcati and his group for their warm hospitality. Some ideas for this paper arose when Pan visited the Istituto per le Applicazioni del Calcolo of CNR and the Department of Mathematics of the University of Rome "La Sapienza"; he wishes to thank professor Roberto Natalini for his kind hospitality.

REFERENCES

[1] C. M. Dafermos, *A system of hyperbolic conservation laws with frictional damping*, Z. Angew. Math. Phys., 46 (1995), pp. 294–307.

[2] C. J. Van Duyn and L. A. Peletier, *Asymptotic behavior of solutions of a nonlinear diffusion equations*, Arch. Rational Mech. Anal., 65 (1977), pp. 363–377.

[3] C. J. Van Duyn and L. A. Peletier, *A class of similary solutions of the nonlinear diffusion equations*, Nonlinear Anal., 1 (1977), pp. 223–233.

[4] L. Hsiao, *Quasilinear Hyperbolic Systems and Dissipative Mechanisms*, World Scientific, River Edge, NJ, 1997.

[5] L. Hsiao and T. P. Liu, *Convergence to nonlinear diffusion waves for solutions of a system of hyperbolic conservation laws with damping*, Comm. Math. Phys., 143 (1992), pp. 599–605.

[6] L. Hsiao and T. P. Liu, *Nonlinear diffusive phenomena of nonlinear hyperbolic systems*, Chinese Ann. Math. Ser B, 14 (1993), pp. 465–480.

[7] L. Hsiao and T. Luo, *Nonlinear diffusive phenomena of solutions for the system of compressible adiabatic flow through porous media*, J. Differential Equations, 125 (1996), pp. 329–365.

[8] L. Hsiao and T. Luo , *Nonlinear diffusive phenomena of entropy weak solutions for a system of quasilinear hyperbolic conservation laws with damping*, Quart. Appl. Math., 56 (1998), pp. 173–198.

[9] L. Hsiao and R. H. Pan, *Initial boundary value problem for the system of compressible adiabatic flow through porous media*, J. Differential Equations, 159 (1999), pp. 280–305.

[10] L. Hsiao and D. Serre, *Global existence of solutions for the system of compressible adiabatic flow through porous media*, SIAM J. Math. Anal., 27 (1996), pp. 70–77.

[11] L. Hsiao and D. Serre, *Large-time behavior of solutions for the system of comprssible adiabatic flow through porous media*, Chinese Ann. Math. Ser. B, 16 (1995), pp. 1–14.

[12] L. Hsiao and S. Q. Tang, *Construction and qualitative behavior of solutions for a system of nonlinear hyperbolic conservation laws with damping*, Quart. Appl. Math., 53 (1995), pp. 487–505.

[13] L. Hsiao and S. Q. Tang, *Construction and qualitative behavior of the solution of the perturbated Riemann problem for the system of one-dimensional isentropic flow with damping*, J. Differential Equations, 123 (1995), pp. 480–503.

[14] T. Luo and T. Yang, *Interaction of elementary waves for compressible Euler equations with frictional damping*, J. Differential Equations, 161 (2000), pp. 42–86.

[15] P. Marcati and A. Milani, *The one dimensional Darcy's law as the limit of a compressible Euler flow*, J. Differential Equations, 84 (1990), pp. 129–147.

[16] P. Marcati and M. Mei, *Convergence to nonlinear diffusion waves for solutions of the initial boundary problem to the hyperbolic conservation laws with damping*, Quart. Appl. Math., 53 (2000), pp. 763–784.

[17] P. Marcati, A. Milani, and P. Secchi, *Singular convergence of weak solutions for a quasilinear nonhomogeneous hyperbolic system*, Manuscripta Math., 60 (1988), pp. 49–69.

[18] P. Marcati and B. Rubino, *Hyperbolic to parabolic relaxation theory for quasilinear first order systems*, J. Differential Equations, 162 (2000), pp. 359–399.

[19] K. Nishihara, *Convergence rates to nonlinear diffusion waves for solutions of system of hyperbolic conservation laws with damping*, J. Differential Equations, 131 (1996), pp. 171–188.

[20] K. Nishihara and T. Yang, *Boundary effect on asymptotic behavior of solutions to the p-system with damping*, J. Differential Equations, 161 (2000), pp. 191–218.

[21] D. Serre and L. Xiao, *Asymptotic behavior of large weak entropy solutions of the damped p-system*, J. Partial Differential Equations, 10 (1997), pp. 355–368.

[22] Y. S. Zheng, *Global smooth solutions to the adiabatic gas dynamics system with dissipation terms*, Chinese J. Contemp. Math., 17 (1996), pp. 155–162.

SIAM J. Math. Anal.
Vol. 33, No. 4, pp. 827–846

A REPRESENTATION FORMULA FOR THE MEAN CURVATURE MOTION*

R. BUCKDAHN[†], P. CARDALIAGUET[†], AND M. QUINCAMPOIX[†]

Abstract. The goal of this paper is to give a representation formula for the mean curvature motion in terms of the value function of some stochastic optimal control problem. This result is generalized to several geometric evolution equations.

Key words. stochastic optimal control, viability, viscosity solutions

AMS subject classifications. 53C44, 35K65, 49L25, 93E03, 60H10, 60H30

PII. S0036141000380334

Introduction. The main result of this paper is a representation formula for the solutions of the equation of the mean curvature motion (MCM). This equation is the following second order parabolic (degenerate and nonlinear) PDE:

$$\left\{ \begin{array}{ll} u_t = |Du| curv(u) & \text{on } (0,+\infty)\times\mathbb{R}^n, \\ u(0,\cdot) = g & \text{on } \mathbb{R}^n, \end{array} \right. \tag{1}$$

where $curv(u) = \text{div}(Du/|Du|) = (\Delta u - \frac{\langle D^2uDu, Du\rangle}{|Du|^2})/|Du|$. It is known (cf. [5], [7]) that this equation has a unique solution in the viscosity sense when $g : \mathbb{R}^n \to \mathbb{R}^n$ is bounded and uniformly continuous. We have obtained the following formula: For any $t > 0$,

$$u(t,x) = \inf_{v\in\mathcal{A}} \left(\text{ess-sup}_\Omega\, g(X^{x,v(\cdot)}(t))\right),$$

where "$\inf_{v\in\mathcal{A}}$" means the infimum over any complete stochastic basis $(\Omega, \mathcal{F}, P; (\mathcal{F}_s, s \in [0,T]))$ endowed with an n-dimensional standard $(\mathcal{F}_s)$-Brownian motion $W = (W(s),\, s \in [0,T])$, and over any $(\mathcal{F}_s)$-progressively measurable process $v(\cdot)$ taking its values in the set

$$\mathcal{V} = \{v \in \mathcal{S}_n \mid v \geq 0,\ I - v^2 \geq 0, \text{ and } \text{Tr}(I - v^2) = 1\}$$

(in brief: $v \in \mathcal{A}(\Omega, \mathcal{F}, P; W)$). The process $X^{x,v(\cdot)} = (X^{x,v(\cdot)}(s),\, s \in [0,T])$ is the solution of the associated stochastic control system

$$\left\{ \begin{array}{l} dX^{x,v(\cdot)}(s) = \sqrt{2}v(s)dW(s), \\ X^{x,v(\cdot)}(0) = x. \end{array} \right. \tag{2}$$

As a byproduct of this result we obtain a representation formula for the MCM. A family of moving hypersurfaces (Σ_t) of $\mathbb{R}^n$ (without boundary) is evolving by its mean curvature if its normal velocity is equal at each point to its mean curvature at that point. It is known that, in general, there is no regular solution of the MCM: the

[*] Received by the editors November 1, 2000; accepted for publication (in revised form) August 22, 2001; published electronically December 18, 2001.

http://www.siam.org/journals/sima/33-4/38033.html

[†] Département de Mathématiques, Université de Bretagne Occidentale, 6, avenue Victor-le-Gorgeu, B.P. 809, 29285 Brest cedex, France (Rainer.Buckdahn@univ-brest.fr, Pierre.Cardaliaguet@univ-brest.fr, Marc.Quincampoix@univ-brest.fr).

family of moving surfaces (Σ_t) usually develops singularities in finite time. Several definitions for the front after the apparitions of singularities have been proposed. Here we use, on the one hand, the level set method introduced in [5] and [7], and, on the other hand, the distance solutions of a front introduced in an equivalent way in [11] and [3].

In the level set method, the front (Σ_t) is defined as the 0-level set of the solution u of (1) for some uniformly continuous function g vanishing on Σ_0. The level set method gives a unique solution for the front (independent of g) but has to face the fattening problem: The set Σ_t may have a nonempty interior.

A distance solution of the MCM is a family of moving sets (Σ_t) such that the function $u(t,x) = \mathbf{1}_{\Sigma_t}(x)$ is a (discontinuous) solution to (1) with $g = \mathbf{1}_{\Sigma_0}(x)$. Compared with the level set method, this way of defining a solution is more intrinsic. Unfortunately, distance solutions are not unique in general. In fact, the fattening problem in the level method and the nonuniqueness problem of distance solutions are directly related: Indeed, the solution given by the level set method is always a distance solution, and fattening occurs if and only if this solution is not the unique one. The solution given by the level set method is sometimes called the biggest flow because it contains any distance solution.

Let us finally recall that if there is a smooth solution to the MCM on the interval $[0,T]$, then this solution is unique and the distance solutions—and therefore also the biggest flow—coincide with the smooth solution on $[0,T]$.

We obtain the following representation formula for the MCM: Let (Σ_t) be the biggest flow of the MCM. Then we have

$$\forall t \geq 0,\ \ \Sigma_t = \{x \in \mathbb{R}^n \mid \exists v(\cdot) \in \mathcal{A} \text{ such that } X^{x,v(\cdot)}(t) \in \Sigma_0 \text{ a.s.}\}, \tag{3}$$

where "$\exists v(\cdot) \in \mathcal{A}$" means that there is a complete stochastic basis $(\Omega, \mathcal{F}, P; (\mathcal{F}_s,\ s \in [0,T]))$ endowed with an n-dimensional standard $(\mathcal{F}_s)$-Brownian motion $W = (W(s),\ s \in [0,T])$ and that $v(\cdot)$ is some $\mathcal{V}$-valued and $(\mathcal{F}_s)$-progressively measurable process. The process $X^{x,v(\cdot)}$ is the solution to (2).

We actually prove that (Σ_t) satisfies a stronger property:

$$\forall t \geq 0,\ \ x \in \Sigma_t \ \Leftrightarrow\ \exists v(\cdot) \in \mathcal{A} \text{ such that } X^{x,v(\cdot)}(s) \in \Sigma_s \text{ for } s \in [0,t]. \tag{4}$$

This statement can be understood as a dynamic programming principle. In fact, (4) not only holds true for the solution of the MCM given by the level set method but also for the distance solutions of the MCM. This property even characterizes these distance solutions. Indeed, if (Σ_t) is a family of moving sets, we prove that (Σ_t) is a distance solution of the MCM if and only if both (Σ_t) and $(\mathbb{R}^n \backslash \Sigma_t)$ satisfy a dynamic programming principle similar to (4).

This characterization of distance solutions relies upon a viability theorem for moving sets. Similar viability results can be found in the literature: see, in particular, [1], [2], [4], [9].

Let us now explain how this paper is organized. In section 1, we establish the representation formula for the solution of (1). Section 2 is devoted to the solutions of the MCM given by the level set approach, while the study of the distance solutions is the aim of section 3. In section 4, we generalize these results to some more general geometric equations. We complete the paper by giving in the appendix a statement of the viability theorem for moving sets and its proof.

When this paper was almost complete, we learned that Soner and Touzi [12] had a representation result similar to ours. In fact, their result is applicable to the motion by mean curvature in any codimension. In codimension one (i.e., the case we consider here), they prove that the moving set (Σ_t) given by

$$\Sigma_t = \{x \in \mathbb{R}^n \mid \exists v(\cdot) \in \mathcal{A} \text{ such that } X^{x,v(\cdot)}(t) \in \Sigma_0 \text{ a.s.}\}$$

is a distance solution of the MCM. Our result is slightly more precise, since it shows that (Σ_t) is in fact the solution given by the level set method.

1. A control problem.

Let us first define the set of controls: Let $\mathcal{S}_n$ be the set of all $n \times n$ symmetric matrices, and let $\mathcal{V}$ be the following compact subset of $\mathcal{S}_n$:

$$\mathcal{V} = \{v \in \mathcal{S}_n \mid v \geq 0,\ I - v^2 \geq 0 \text{ and } \mathrm{Tr}(I - v^2) = 1\},$$

where I is the $n \times n$ identity matrix. Let us point out that we have the following equality:

$$\{v^2 \mid v \in \mathcal{V}\} = \{w \in \mathcal{S}_n \mid w \geq 0,\ I - w \geq 0,\ \mathrm{Tr}(I - w) = 1\}, \tag{5}$$

from which we deduce that $\{v^2 \mid v \in \mathcal{V}\}$ is a convex subset of $\mathcal{S}_n$ and that

$$\{v^2 \mid v \in \mathcal{V}\} = \mathrm{Co}\,\{(I - aa^*) \mid a \in \mathbb{R}^n,\ |a| = 1\}, \tag{6}$$

where $\mathrm{Co}(A)$ stands for the closed convex hull of a set A.

Let $W = (W(s),\ s \in [0,T])$ be an n-dimensional standard $(\mathcal{F}_s)$-Brownian motion on some complete stochastic basis $(\Omega, \mathcal{F}, P; (\mathcal{F}_s))$. We denote by $\mathcal{A} = \mathcal{A}(\Omega, \mathcal{F}, P; W)$ the set of all $\mathcal{V}$-valued $(\mathcal{F}_s)$-progressively measurable processes $v(\cdot)$. A process $v(\cdot) \in \mathcal{A}$ is called an admissible control.

Let us recall that, if $(\Omega, \mathcal{A}, P)$ is a probability space and if $Y : \Omega \to \mathbb{R}$ is a random variable, then the ess-sup of Y is defined by

$$\text{ess-sup}_\Omega\, Y = \sup\{\tau \in \mathbb{R} \mid P(Y \geq \tau) > 0\}.$$

Our aim is to prove the following result.

THEOREM 1.1. *Let $g : \mathbb{R}^n \to \mathbb{R}$ be a bounded uniformly continuous function. Let $T > 0$ be fixed, and let us set, for any initial position $(t,x) \in [0,T] \times \mathbb{R}^n$,*

$$V(t,x) = \inf_{v \in \mathcal{A}} \left(\text{ess-sup}_\Omega\, g(X^{t,x,v(\cdot)}(T)) \right), \tag{7}$$

where $X^{t,x,v(\cdot)}(\cdot)$ is the solution to

$$\left\{ \begin{array}{rcl} dX^{t,x,v(\cdot)}(s) & = & \sqrt{2} v(s) dW(s)\ , \\ X^{t,x,v(\cdot)}(t) & = & x. \end{array} \right. \tag{8}$$

Then V is the solution, in the viscosity sense, of the equation of the MCM (written here with a terminal condition):

$$\left\{ \begin{array}{ll} -V_t - \Delta V + \frac{\langle D^2 V DV, DV \rangle}{|DV|^2} = 0 & \text{in } (0,T) \times \mathbb{R}^n, \\ V(T,\cdot) = g(\cdot) & \text{in } \mathbb{R}^n. \end{array} \right. \tag{9}$$

Remarks.

1. In particular, the value function V, as the solution to (9), is continuous.

2. We prove below that, for any $(t,x) \in [0,T] \times \mathbb{R}^n$, there is an optimal control $v(\cdot) \in \mathcal{A}$ in the following sense: There is a complete stochastic basis $(\Omega, \mathcal{F}, P; (\mathcal{F}_s,\, s \in [0,T]))$ endowed with an n-dimensional standard $(\mathcal{F}_s)$-Brownian motion $W = (W(s),\, s \in [0,T])$, and some $(\mathcal{F}_s)$-progressively measurable process $v(\cdot)$ taking its values in the set $\mathcal{V}$, such that the solution $X^{t,x,v(\cdot)}$ of (8) satisfies

$$V(t,x) = \text{ess-sup}_\Omega\, g(X^{t,x,v(\cdot)}(T)).$$

Proof of Theorem 1.1. Since, for any constants $a > 0$ and $b \in \mathbb{R}$, the function $(t,x) \to aV(t,x) + b$ is the value function for the ess-sup equation (7) for the terminal cost $x \to ag(x) + b$, and since g is bounded, we can assume, without loss of generality, that

$$\forall x \in \mathbb{R}^n,\ 1 \leq g(x) \leq 2. \tag{10}$$

Let us set, for any $p \geq 1$ and any $(t,x) \in [0,T] \times \mathbb{R}^n$,

$$V_p(t,x) = \inf_{v \in \mathcal{A}} \left[E((g(X^{t,x,v(\cdot)}(T)))^p) \right]^{\frac{1}{p}} ,$$

where $X^{t,x,v(\cdot)}(\cdot)$ is the solution to (8). It is known (see [8]) that V_p^p is the solution, in the viscosity sense, of the following equation:

$$\left\{ \begin{array}{ll} -(V_p^p)_t + \mathcal{H}(D^2 V_p^p) = 0 & \text{in } (0,T) \times \mathbb{R}^n, \\ V_p^p(T,\cdot) = g^p(\cdot) & \text{in } \mathbb{R}^n, \end{array} \right. \tag{11}$$

where

$$\forall S \in \mathcal{S}_n,\ \mathcal{H}(S) = \sup_{v \in \mathcal{V}} \left[-\text{Tr}(vv^* S) \right].$$

Since, from (6), the elements of $\mathcal{V}$ are symmetric matrices v such that v^2 are convex combinations of matrices of the form $(I - aa^*)$, we get

$$\forall S \in \mathcal{S}_n,\ \mathcal{H}(S) = \sup_{|a|=1} \left[-\text{Tr}((I - aa^*)S) \right] = - \left[\text{Tr}(S) - \lambda_{max}(S) \right],$$

where $\lambda_{max}(S)$ is the largest eigenvalue of S because

$$\max_{|a|=1} \text{Tr}(aa^* S) = \max_{|a|=1} \langle Sa, a \rangle = \lambda_{max}(S).$$

We divide the proof of Theorem 1.1 in two steps. In the first step, we prove that (V_p) converges to V. In the second step, we deduce from (11), satisfied by the V_p^p, that V is the solution of (9).

First step. Let $(t,x) \in [0,T] \times \mathbb{R}^n$ be fixed. We claim that

$$\lim_{p \to +\infty} V_p(t,x) = V(t,x). \tag{12}$$

Proof of the first step. Since, for any $v(\cdot) \in \mathcal{A}$ and any $1 \leq p \leq q$, we have

$$\left[E((g(X^{t,x,v(\cdot)}(T)))^p) \right]^{\frac{1}{p}} \leq \left[E((g(X^{t,x,v(\cdot)}(T)))^q) \right]^{\frac{1}{q}} \leq \text{ess-sup}_\Omega\, g(X^{t,x,v(\cdot)}(T)),$$

we deduce that the sequence $(V_p(t,x))$ is nondecreasing and that

$$\lim_{p\to+\infty} V_p(t,x) \le V(t,x).$$

Let us now prove the converse inequality. For doing so, we consider, for any $p \ge 1$, a control $v_p(\cdot)$ such that

$$\Big[E((g(X^{t,x,v_p(\cdot)}(T)))^p)\Big]^{\frac{1}{p}} \le V_p(t,x) + \frac{1}{p}.$$

Recall that $\mathcal{V}$ is compact. From (6), the set $\{\frac{1}{2}(\sqrt{2}v)(\sqrt{2}v)^* \mid v \in \mathcal{V}\}$ is convex. Hence, from standard arguments,[1] there exist a probability space $(\Omega, \mathcal{F}, P)$, some n-dimensional Brownian motion W on this space, some process $v(\cdot) \in \mathcal{A}(\Omega, \mathcal{F}, P; W)$, and a subsequence p_k such that

$$\forall q \ge 1,\ \lim_{k\to+\infty} \Big[E((g(X^{t,x,v_{p_k}(\cdot)}(T)))^q)\Big]^{\frac{1}{q}} = \Big[E((g(X^{t,x,v(\cdot)}(T)))^q)\Big]^{\frac{1}{q}}.$$

We have, for any fixed $q \ge 1$ and k sufficiently large,

$$\Big[E((g(X^{t,x,v_{p_k}(\cdot)}(T)))^q)\Big]^{\frac{1}{q}} \le \Big[E((g(X^{t,x,v_{p_k}(\cdot)}(T)))^{p_k})\Big]^{\frac{1}{p_k}} \le V_{p_k}(t,x) + \frac{1}{p_k}.$$

Since the sequence $(V_p(t,x))$ has a limit when $p \to +\infty$, letting $k \to +\infty$, we get

$$\Big[E((g(X^{t,x,v(\cdot)}(T)))^q)\Big]^{\frac{1}{q}} \le \lim_{p\to+\infty} V_p(t,x).$$

Then, letting $q \to +\infty$, we obtain the desired result:

$$V(t,x) \le \text{ess-sup}_{\Omega}\, g(X^{t,x,v(\cdot)}(T)) \le \lim_{p\to+\infty} V_p(t,x).$$

In particular, we see that the admissible control $v(\cdot) \in \mathcal{A}$ is optimal.

Second step. We now prove that V is the solution of (9). For doing so, we use the fact that V can be approximated by the V_p. Let us point out that V is lower semicontinuous, as the supremum of the continuous maps V_p. We do not prove directly that V is a solution. Instead, we consider the half-relaxed upper-limit $V^\sharp$ of the V_p:

$$\forall (t,x) \in [0,T]\times\mathbb{R}^n,\ V^\sharp(t,x) = \limsup_{(t',x')\to(t,x),\ p\to+\infty} V_p(t',x').$$

We are going to prove that V is a supersolution and that $V^\sharp$ is a subsolution. Moreover, V and $V^\sharp$ satisfy the following boundary conditions: $V(T,\cdot) \ge g$ and $V^\sharp(T,\cdot) \le g$. Then we obtain the equality $V = V^\sharp$ by the comparison principle since clearly

$$\forall (t,x),\ V(t,x) \le V^\sharp(t,x).$$

Equation satisfied by $\boldsymbol{V_p}$. From (11), one can deduce easily that V_p is a solution to

$$\begin{cases} -pV_p^{p-1}(V_p)_t + \mathcal{H}(p(p-1)V_p^{p-2}DV_p(DV_p)^* + pV_p^{p-1}D^2V_p) = 0 & \text{in } (0,T)\times\mathbb{R}^n, \\ V_p(T,\cdot) = g(\cdot) & \text{in } \mathbb{R}^n. \end{cases}$$

[1] See, in particular, [6] or [13, Theorem 5.3, Chapter 2] and its proof, namely, relations (5.25) and (5.26) and the conclusion from (5.41) to (5.43).

Since, from assumption (10), we have $g \geq 1$, we also have $V_p \geq 1$. Thus we can divide the above equation by pV_p^{p-1} to get that V_p is in fact the solution to

$$(13) \qquad \left\{ \begin{array}{ll} -(V_p)_t + \mathcal{H}((p-1)V_p^{-1}DV_p(DV_p)^* + D^2V_p) = 0 & \text{in } (0,T) \times \mathbb{R}^n, \\ V_p(T,\cdot) = g(\cdot) & \text{in } \mathbb{R}^n \end{array} \right.$$

because $\mathcal{H}$ is positively homogeneous of degree one.

***V* is a supersolution**. We now prove that V is a supersolution of (9). Let ϕ be a test function such that $V - \phi$ has a strict local minimum at some point $(t,x) \in (0,T) \times \mathbb{R}^n$. We have to prove that, at the point (t,x), we have

$$-\phi_t + H^*(D\phi, D^2\phi) \geq 0,$$

where we have set

$$\forall d \in \mathbb{R}^n \backslash \{0\},\ \forall S \in \mathcal{S}_n,\ H(d,S) = -\mathrm{Tr}(S) + \frac{\langle Sd, d\rangle}{|d|^2}$$

and where H^* is an upper regularization of H when $d = 0$, namely,

$$\forall S \in \mathcal{S}_n,\ H^*(0,S) = -\mathrm{Tr}(S) + \lambda_{max}(S).$$

Standard arguments show that there is a sequence (t_p, x_p) converging to (t,x) such that $(V_p(t_p,x_p))$ converges to $V(t,x)$ and such that $V_p - \phi$ has a local minimum at (t_p,x_p). Therefore, we have, at the point (t_p,x_p),

$$(14) \qquad -\phi_t + \mathcal{H}((p-1)V_p^{-1}D\phi(D\phi)^* + D^2\phi) \geq 0$$

because V_p is a viscosity solution of (13).

First case. $D\phi(t,x) \neq 0$. In this case, we have to prove that, at the point (t,x), we have

$$-\phi_t - \Delta\phi + \frac{\langle D^2\phi D\phi, D\phi\rangle}{|D\phi|^2} \geq 0.$$

Let us compute $\mathcal{H}((p-1)V_p^{-1}D\phi(D\phi)^* + D^2\phi)$ at the point (t_p,x_p):

$$(15) \qquad \begin{array}{l} \mathcal{H}((p-1)V_p^{-1}D\phi(D\phi)^* + D^2\phi) \\ \quad = -(p-1)V_p^{-1}|D\phi|^2 - \Delta\phi + \lambda_{max}((p-1)V_p^{-1}D\phi(D\phi)^* + D^2\phi) \end{array}$$

from the definition of $\mathcal{H}$. At this stage, we need to recall the following classical result.

LEMMA 1.2. *Let $S \in \mathcal{S}_n$ be such that the space of eigenvectors associated with $\lambda_{max}(S)$ is of dimension one. Then, the map $S \to \lambda_{max}(S)$ is of class $\mathcal{C}^1$ in a neighborhood of S. Moreover, its derivative $D\lambda_{max}(S)$ at this point S is given by*

$$(16) \qquad \forall H \in \mathcal{S}_n,\ D\lambda_{max}(S)(H) = \langle Ha, a\rangle,$$

where $a \in \mathbb{R}^n$ is an eigenvector of S associated with $\lambda_{max}(S)$ and such that $|a| = 1$.

(The proof of the lemma is given at the end of the present section for the sake of completeness.)

We apply the lemma to the matrix

$$S = \frac{D\phi(t,x)(D\phi(t,x))^*}{V(t,x)}$$

for which $\lambda_{max}(S) = \frac{|D\phi(t,x)|^2}{V(t,x)}$, and where $a = D\phi(t,x)/|D\phi(t,x)|$ since $D\phi(t,x) \neq 0$. Let us also set

$$S_p = \frac{D\phi(t_p,x_p)(D\phi(t_p,x_p))^*}{V_p(t_p,x_p)},$$

and let us notice that S_p converges to S. Moreover, since the matrices S_p satisfy the conditions of Lemma 1.2, the Taylor formula states that there is some $\theta_p \in (0,1)$ such that

$$\lambda_{max}\left(S_p + \frac{D^2\phi_p}{p-1}\right) = \lambda_{max}(S_p) + \frac{1}{p-1} D\lambda_{max}\left(S_p + \frac{\theta_p}{p-1} D^2\phi_p\right)(D^2\phi_p),$$

where we have set, for simplicity, $D\phi_p = D\phi(t_p,x_p)$ and $D^2\phi_p = D^2\phi(t_p,x_p)$. We now use the fact that λ_{max} is $\mathcal{C}^1$ in a neighborhood of S and that $S_p \to S$ to get

$$\lambda_{max}\left(S_p + \frac{D^2\phi_p}{p-1}\right) = \lambda_{max}(S_p) + \frac{1}{p-1} D\lambda_{max}(S)(D^2\phi_p) + o(1/p),$$

where $p\,o(1/p) \to 0$ when $p \to +\infty$. Then formula (16) gives

$$\lambda_{max}\left(S_p + \frac{D^2\phi_p}{p-1}\right) = \lambda_{max}(S_p) + \frac{\langle D^2\phi_p D\phi(t,x), D\phi(t,x)\rangle}{(p-1)|D\phi(t,x)|^2} + o(1/p).$$

Using this last equality in (15), we get, at the point (t_p,x_p),

$$\mathcal{H}((p-1)V_p^{-1}D\phi(D\phi)^* + D^2\phi) = -\Delta\phi + \frac{\langle D^2\phi D\phi(t,x), D\phi(t,x)\rangle}{|D\phi(t,x)|^2} + p\,o(1/p),$$

which, combined with (14), gives the desired inequality when $p \to +\infty$:

$$-\phi_t - \Delta\phi + \frac{\langle D^2\phi D\phi, D\phi\rangle}{|D\phi|^2} \geq 0$$

at the point (t,x).

Second case. $D\phi(t,x) = 0$. In this case, we have to prove that, at the point (t,x), we have

$$-\phi_t - \Delta\phi + \lambda_{max}(D^2\phi) \geq 0.$$

In order to use inequality (14), we first compute the quantity

$$\mathcal{H}((p-1)V_p^{-1}D\phi(D\phi)^* + D^2\phi)$$

taken at the point (t_p,x_p). Let us notice that, since $S \to \lambda_{max}(S)$ is subadditive, we have

$$\lambda_{max}((p-1)V_p^{-1}D\phi(D\phi)^* + D^2\phi) \leq (p-1)V_p^{-1}|D\phi|^2 + \lambda_{max}(D^2\phi),$$

from which we deduce the following inequality, taken at the point (t_p,x_p):

$$\mathcal{H}((p-1)V_p^{-1}D\phi(D\phi)^* + D^2\phi) \leq -\Delta\phi + \lambda_{max}(D^2\phi).$$

Using this inequality in (14) and letting $p \to +\infty$ give the desired result:

$$-\phi_t - \Delta\phi + \lambda_{max}(D^2\phi) \geq 0.$$

***$V^\sharp$* is a subsolution.** We set, for any positive z, any $d \in \mathbb{R}\backslash\{0\}$, and $S \in \mathcal{S}_n$,

$$\mathcal{H}_p(z,d,S) = \mathcal{H}\left(\frac{p-1}{z}dd^* + S\right).$$

Let us notice that we have

$$\mathcal{H}_p(z,d,S) \geq -\mathrm{Tr}(S) + \frac{\langle Sd, d\rangle}{|d|^2}$$

because

$$\mathcal{H}_p(z,d,S) = -\frac{(p-1)}{z}|d|^2 - \mathrm{Tr}(S) + \lambda_{max}\left(\frac{(p-1)}{z}dd^* + S\right)$$

with

$$\lambda_{max}\left(\frac{(p-1)}{z}dd^* + S\right) \geq \frac{(p-1)}{z}|d|^2 + \frac{\langle Sd, d\rangle}{|d|^2}.$$

Thus the V_p, being the solutions of (13), are subsolutions of (9). Then, from standard argument, the half-relaxed upper-limit of the V_p, which is nothing but $V^\sharp$, is still a subsolution of (9).

***V* satisfies the terminal condition.** Since the V_p satisfy $V_p(T,\cdot) = g$ and since $V_p(T,\cdot) \to V(T,\cdot)$, we clearly have $V(T,\cdot) = g$.

***$V^\sharp$* satisfies the terminal condition.** We have to show that $V^\sharp(T,\cdot) \leq g$. We argue by contradiction, by assuming that there are some $x_0 \in \mathbb{R}^n$ and some $\epsilon > 0$ such that $V^\sharp(T,x_0) \geq g(x_0) + \epsilon$. Let $\beta > 0$ be sufficiently large such that

$$\forall x \in \mathbb{R}^n,\ \frac{\beta}{2}|x-x_0|^2 + g(x_0) + \frac{\epsilon}{2} \geq g(x) + \frac{\epsilon}{4}.$$

(There is such a β because g is bounded and continuous.) We now choose $\alpha > \beta(n-1)$ sufficiently large so that

$$\forall x \in \mathbb{R}^n,\ \alpha T + \frac{\beta}{2}|x-x_0|^2 + g(x_0) + \frac{\epsilon}{2} \geq 3.$$

Let us set

$$\phi(t,x) = \alpha(T-t) + \frac{\beta}{2}|x-x_0|^2 + g(x_0) + \frac{\epsilon}{2}.$$

From the construction of ϕ, we have

$$\forall x \in \mathbb{R}^n,\ \phi(0,x) \geq 3 \text{ and } \phi(T,x) \geq g(x) + \frac{\epsilon}{4}.$$

Hence we have, for any $p \geq 1$,

$$\forall x \in \mathbb{R}^n,\ \phi(0,x) \geq V_p(0,x) + 1 \text{ and } \phi(T,x) \geq V_p(T,x) + \frac{\epsilon}{4} \tag{17}$$

because $1 \leq V_p \leq 2$ and $V_p(T,\cdot) = g$. From our assumption, $V^\sharp(T,x_0) \geq g(x_0) + \epsilon$, we deduce that there are a subsequence p_k and some (t_k, x_k) converging to (T, x_0) such that

$$\lim_{k\to+\infty} V_{p_k}(t_k,x_k) = V^\sharp(T,x_0) \geq g(x_0) + \epsilon.$$

Hence, for k sufficiently large, we have

$$V_{p_k}(t_k, x_k) > \phi(t_k, x_k) \tag{18}$$

since $\phi(T, x_0) = g(x_0) + \epsilon/2$. Let us now notice that $V_{p_k} - \phi$ has a local maximum at some point $(s_k, y_k) \in [0, T] \times \mathbb{R}^n$ because $-(V_{p_k} - \phi)$ is coercive. From (18), this maximum is positive. Thus, from (17), we have $s_k \in (0, T)$. Since V_{p_k} is the solution to (13), we have

$$\alpha + \mathcal{H}((p_k - 1)V_{p_k}^{-1}(s_k, y_k)\beta^2(y_k - x_0)(y_k - x_0)^* + \beta I) \leq 0, \tag{19}$$

where

$$\mathcal{H}((p_k - 1)V_{p_k}^{-1}(s_k, y_k)\beta^2(y_k - x_0)(y_k - x_0)^* + \beta I) = -\beta(n - 1).$$

This is in contradiction with (19) since $\alpha > \beta(n - 1)$. Therefore, we have proved that $V^\sharp(T, \cdot) \leq g$.

Conclusion. From the previous steps, we know that $V^\sharp$ is a subsolution of (9) with $V^\sharp(T, \cdot) \leq g$, while V is a lower semicontinous supersolution of (9) with $V(T, \cdot) \geq g$. Let us also recall that

$$\forall (t, x) \in [0, T] \times \mathbb{R}^n, \ V(t, x) \leq V^\sharp(t, x).$$

Therefore,

$$\forall x \in \mathbb{R}^n, \ V(T, x) = V^\sharp(T, x) = g(x).$$

Since g is uniformly continuous and since V and $V^\sharp$ are globally bounded, the comparison principle of [10] states that $V^\sharp \leq V$. Therefore, we have proved that $V^\sharp = V$ is the solution to (9).

Remark. In particular, V is continuous.

Proof of Lemma 1.2. Let us recall that

$$\forall S \in \mathcal{S}_n, \quad \lambda_{max}(S) = \max_{|a|=1} \langle Sa, a \rangle.$$

Hence, if $\dim(\mathrm{Ker}(S - \lambda_{max}(S)I)) = 1$, standard argument show that λ_{max} is differentiable at S and that

$$\forall H \in \mathcal{S}_n, \quad D\lambda_{max}(S)(H) = \langle Ha(S), a(S) \rangle,$$

where $a(S)$ is an eigenvalue associated with $\lambda_{max}(S)$ with $|a(S)| = 1$. Moreover, the map $(\lambda, S') \to \dim(\mathrm{Ker}(S' - \lambda I))$ being upper semicontinuous, there is a neighborhood $\mathcal{O}$ of S on which $\dim(\mathrm{Ker}(S' - \lambda_{max}(S')I)) = 1$ for any $S' \in \mathcal{O}$. Hence λ_{max} is differentiable at S' for $S' \in \mathcal{O}$, and its derivative is again given by $D\lambda_{max}(S')(H) = \langle Ha(S'), a(S') \rangle$. This map is clearly continuous (because $\dim(\mathrm{Ker}(S' - \lambda_{max}(S')I)) = 1$), and we have proved that λ_{max} is $\mathcal{C}^1$ in a neighborhood of S. □

2. The representation formula. In this section, we briefly recall the level set method for the MCM, and we prove the representation formula given in the introduction.

Let Σ_0 be a closed subset of $\mathbb{R}^n$, and let $g : \mathbb{R}^n \to \mathbb{R}$ be a uniformly continuous bounded function vanishing on Σ_0:

$$\Sigma_0 = \{x \in \mathbb{R}^n \mid g(x) = 0\}. \tag{20}$$

Let us consider the solution (in the viscosity sense) to

$$\left\{ \begin{array}{ll} u_t = |Du| curv(u) & \text{on } (0,+\infty)\times\mathbb{R}^n, \\ u(0,\cdot) = g & \text{on } \mathbb{R}^n, \end{array} \right. \tag{21}$$

where $curv(u) = \mathrm{div}(Du/|Du|) = (\Delta u - \frac{\langle D^2uDu, Du\rangle}{|Du|^2})/|Du|$. It is known (see [5], [7]) that the set

$$\Sigma_t = \{x \in \mathbb{R}^n \mid u(t,x) = 0\}$$

does not depend on g but only on $\Sigma_0 = \{g = 0\}$: Namely, if g_1 and g_2 are two uniformly continuous, bounded functions, such that

$$\Sigma_0 = \{g_1 = 0\} = \{g_2 = 0\},$$

then the associated solutions u_1 and u_2 satisfy

$$\Sigma_t = \{u_1(t,\cdot) = 0\} = \{u_2(t,\cdot) = 0\}.$$

The moving sets (Σ_t) can be understood as a generalized motion by mean curvature. Indeed, if there is a classical solution to the MCM starting from Σ_0 on some interval $[0,T]$, then this classical solution coincides with (Σ_t) on $[0,T]$.

Let us recall that the front (Σ_t) built by this method is sometimes called the biggest flow because it contains any distance solution of the MCM. (See the next section for the definition of distance solution.)

Remark. The connection between (9) and (21) is the following: A function $u : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}$ is a solution to (21) if and only if, for any $T > 0$, the function $u^T(s,y) = u(T-s,y)$ is a solution to (9).

THEOREM 2.1. *Let (Σ_t) be defined as above. Then*

$$\Sigma_t = \{x \in \mathbb{R}^n \mid \exists v(\cdot) \in \mathcal{A} \;\; \textit{such that } X^{x,v(\cdot)}(t) \in \Sigma_0 \;\textit{a.s.}\},$$

where $X^{x,v(\cdot)}(s)$ is the solution to

$$\left\{ \begin{array}{l} dX^{x,v(\cdot)}(s) = \sqrt{2}v(s)dW(s), \\ X^{x,v(\cdot)}(0) = x. \end{array} \right. \tag{22}$$

Proof of Theorem 2.1. Let $t > 0$ be fixed. Let g be a nonnegative, bounded, uniformly continuous function satisfying (20), and let u be the solution to (21). Let us set $V(s,y) = u(t-s,y)$. Then V is the solution to (9), with a terminal time $T = t$; hence it is equal to the value function defined by (7).

Therefore, a point x belongs to Σ_t if and only if $V(0,x) = u(t,x) = 0$, i.e., there is some optimal control $v(\cdot) \in \mathcal{A}$ for $V(0,x)$ (cf. the remark to Theorem 1.1) such that the solution $X^{0,x,v(\cdot)}$ of (8) satisfies

$$\text{ess-sup}_\Omega g(X^{0,x,v(\cdot)}(t)) = 0.$$

Since g is nonnegative and satisfies (20), this means that

$$X^{0,x,v(\cdot)}(t) \in \Sigma_0 \quad \text{a.s.}$$

Moreover, $X^{0,x,v(\cdot)}$ of (8) is nothing but the solution $X^{x,v(\cdot)}$ of (22). This shows that a point x belongs to Σ_t if and only if there is some control $v(\cdot) \in \mathcal{A}$ such that $X^{x,v(\cdot)}(t) \in \Sigma_0$ a.s. $\square$

3. Characterization of distance solutions of the MCM. We say that a family (Σ_t) of moving sets is a distance solution of the MCM if the function $u(t,x) = \mathbf{1}_{\Sigma_t}(x)$ is a discontinuous solution of the following equation:

$$\left\{\begin{array}{ll} u_t = |Du|curv(u) & \text{on } (0,+\infty)\times\mathbb{R}^n, \\ u(0,\cdot) = \mathbf{1}_{\Sigma_0} & \text{on } \mathbb{R}^n. \end{array}\right. \tag{23}$$

This means that the half-relaxed upper-limit u^* is a subsolution of (23), with $u^* \leq (\mathbf{1}_{\Sigma_0})^*$ at $t=0$, while the half-relaxed lower-limit u_* is a subsolution of (23), with $u_* \geq (\mathbf{1}_{\Sigma_0})_*$ at $t=0$. Let us point out that distance solutions are not unique in general. The biggest solution, i.e., the solution given by the level set method, is a distance solution.

Notation. If (Σ_t) is a family of moving sets, we set

$$\bar{\Sigma}_t = L_{t'\to t}\Sigma_{t'} = \{x\in\mathbb{R}^n \mid \exists t_k\to t,\ x_k\in\Sigma_{t_k} \text{ with } x_k\to x\}.$$

In the same way, we set

$$\widehat{\Sigma}_t = \mathrm{Limsup}_{t'\to t}\mathbb{R}^n\backslash\Sigma_t = \{x\in\mathbb{R}^n \mid \exists t_k\to t,\ x_k\notin\Sigma_{t_k} \text{ with } x_k\to x\}.$$

We characterize the fact that (Σ_t) is moving by the mean curvature in terms of the viability of (Σ_t) and $(\mathbb{R}^n\backslash\Sigma_t)$ for the stochastic control system (22).

PROPOSITION 3.1. *Let (Σ_t) be a family of moving sets satisfying the initial condition:*

$$\overline{\Sigma_0} = \bar{\Sigma}_0 \quad \text{and} \quad \overline{\mathbb{R}^n\backslash\Sigma_0} = \widehat{\Sigma}_0. \tag{24}$$

Then (Σ_t) is a distance solution of the MCM if and only if the two following properties are satisfied:

(i) *For any $t\geq 0$ and any $x\in\bar{\Sigma}_t$, there exists some $v(\cdot)\in\mathcal{A}$ such that the solution $X^{x,v(\cdot)}$ of* (22) *satisfies*

$$\forall s\in[0,t], \qquad X^{x,v(\cdot)}(s)\in\bar{\Sigma}_{t-s} \quad a.s.$$

(ii) *For any $t\geq 0$ and any $x\in\widehat{\Sigma}_t$, there exists some $v(\cdot)\in\mathcal{A}$ such that the solution $X^{x,v(\cdot)}$ of* (22) *satisfies*

$$\forall s\in[0,t], \qquad X^{x,v(\cdot)}(s)\in\widehat{\Sigma}_{t-s} \quad a.s.$$

Remarks.

1. Actually, property (i) is equivalent with the fact that u is a subsolution of (23), while property (ii) is equivalent with the fact that u is a supersolution.
2. Since the solution given by the level set method is a distance solution, this means that this solution satisfies (i) and (ii).

The proposition is an application of a stochastic viability theorem (Theorem A.1) given in the appendix.

Proof of Proposition 3.1. We are going to prove that the map $u(t,x) = \mathbf{1}_{\Sigma_t}(x)$ is a subsolution of (23) if and only if (Σ_t) satisfies (i). The proof that u is a supersolution of (23) if and only if (Σ_t) satisfies (ii) is similar, and so we omit it.

Let us notice that $u^*(t,x) = \mathbf{1}_{\bar{\Sigma}_t}(x)$. From Theorem A.1, (ii), applied to the control system (8), the fact that $u(t,x) = \mathbf{1}_{\Sigma_t}(x)$ is a subsolution of (23) means that, for any $T>0$, the family of moving sets $(K_t^T)_{t\in[0,T]}$ defined by

$$\forall t\in[0,T],\ \ K_t^T = \bar{\Sigma}_{T-t}$$

is a viability domain (cf. Definition A.2 below) for the stochastic control system (8).

Let us first assume that u is a subsolution of (23). Then, for any $t > 0$ and any $x \in \Sigma_t$, the viability theorem applied to the viability domain $(K_s^T)_{s\in[0,T]}$, for some $T > t$, together with Remark A.1 part 3, implies that there is some control $v \in \mathcal{A}$ such that the solution $X^{0,x,v(\cdot)}$ of (8) satisfies

$$\forall s \in [0,t],\ X^{0,x,v(\cdot)}(s) \in K_{T-t+s}^T = \bar{\Sigma}_{t-s},$$

because the point x belongs to $K_{T-t}^T = \bar{\Sigma}_t$. Therefore, part (i) of the proposition is satisfied since the solution $X^{0,x,v(\cdot)}$ is nothing but the solution $X^{x,v(\cdot)}$ of (22).

Conversely, let us now assume that assertion (i) of the proposition holds true. Then Remark A.1 part 3 implies that, for any $T > 0$, the family (K_t^T) is a viability domain for system (8). Hence u is a subsolution of (23). □

4. Representation formula for anisotropic flows. In this section we consider geometric equations of the form

$$u_t = F(Du, D^2u). \tag{25}$$

By geometric, we mean that the function $F : \mathbb{R}^n\backslash\{0\} \times \mathcal{S}_n \to \mathbb{R}$ is elliptic, i.e.,

$$F(p,A) \le F(p,B) \qquad \text{whenever } A \le B \tag{26}$$

and satisfies the following conditions: for all $a \in \mathbb{R}\backslash\{0\}$ and for all $\sigma \in \mathbb{R}$,

$$F(ap, aX) = aF(p,X) \text{ and } F(p, X + \sigma pp^*) = F(p,X). \tag{27}$$

A geometric equation enjoys the so-called invariance property: If u is a solution to (25) in the viscosity sense, then, for any continuous function $\theta : \mathbb{R} \to \mathbb{R}$, the function $\theta(u)$ also satisfies (25) (cf. [5], [7]). With such an equation, one can associate a geometric flow: If Σ_0 is a closed subset of $\mathbb{R}^n$, the flow (Σ_t) of Σ_0 by F is defined by

$$\forall t \ge 0,\ \ \Sigma_t = \{x \in \mathbb{R}^n \mid u(t,x) = 0\},$$

where $u : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}$ is the viscosity solution to

$$\left\{ \begin{array}{ll} u_t = F(Du, D^2u) & \text{on } (0,+\infty) \times \mathbb{R}^n, \\ u(0,\cdot) = g & \text{on } \mathbb{R}^n, \end{array} \right. \tag{28}$$

and where $g : \mathbb{R}^n \to \mathbb{R}^n$ is some uniformly continuous and bounded function vanishing on Σ_0. If F satisfies (27) and the following regularity conditions:

$$\left\{ \begin{array}{rl} \text{(i)} & F \text{ is continuous on } (\mathbb{R}^n\backslash\{0\}) \times \mathcal{S}_n, \\ \text{(ii)} & F \text{ is bounded on bounded subset of } (\mathbb{R}^n\backslash\{0\}) \times \mathcal{S}_n, \\ \text{(iii)} & F^*(0,0) = F_*(0,0) = 0, \end{array} \right. \tag{29}$$

then the set Σ_t depends only on Σ_0 and not on g.

In this section, we give a representation formula for the solution of (28) and for the flow (Σ_t) when F satisfies (26), (27), (29), and the following additional conditions:

$$\left\{ \begin{array}{rl} \text{(i)} & S \to F(p,S) \text{ is concave for any } p \neq 0, \\ \text{(ii)} & \forall p \neq 0,\ \forall S \in \mathcal{S}_n,\ \forall \lambda \in (\mathbb{R}\backslash\{0\}),\ F(\lambda p, S) = F(p,S), \\ \text{(iii)} & \forall p \neq 0,\ \forall S \in \mathcal{S}_n,\ \forall x \notin \mathbb{R}p,\ F(p, S + xx^*) > F(p,S). \end{array} \right. \tag{30}$$

We comment upon these assumptions after the statement of the lemma.

LEMMA 4.1. *Let* $F : \mathbb{R}^n\backslash\{0\} \times \mathcal{S}_n \to \mathbb{R}$ *satisfy* (26), (27), (29), *and* (30). *There is some compact subset* $\mathcal{V}_F$ *of* $\mathcal{S}_n$ *such that*

$$\forall (p, S) \in \mathbb{R}^n\backslash\{0\} \times \mathcal{S}_n, \;\; F(p, S) = \min_{v\in\mathcal{V}_F,\, vp=0} \mathrm{Tr}(Sv^2), \tag{31}$$

and

$$\{v^2 \mid v \in \mathcal{V}_F\} \quad \textit{is convex and compact.} \tag{32}$$

Remarks.

1. The lemma still holds true if we assume that assumption (27) is satisfied only for $a > 0$. However, in this case, the geometric equation associated with (25) is orientation dependent. This means that the correct definition of the associated flow (Σ_t) by the level set method requires the function g in (28) to be positive in the interior of Σ_0 and negative outside. Then the second part of Theorem 4.2 does not hold true anymore.
2. Conditions (30) (i) and (ii) are clearly necessary for formula (31) to hold true. Although (30) (iii) does not seem necessary, it cannot be omitted. For instance, let us assume that F is of the form
$$F(p, S) = \phi(p/|p|)\left(\mathrm{Tr}(S) - \frac{\langle Sp, p\rangle}{|p|^2}\right),$$
where $\phi : \mathbb{R}^n\backslash\{0\} \to \mathbb{R}$ is a continuous nonnegative function which is homogeneous of degree zero, which vanishes at some point $p_0 \neq 0$ and is not identically zero. Then F satisfies all the assumptions of Lemma 4.1 but (30) (iii). We claim that there is no set $\mathcal{V}_F$ such that equality (31) holds true. Indeed, otherwise, the set $\mathcal{V}_F$ should contain 0 because $F(p_0, \cdot) = 0$. In that case, equality (31) implies that $F \leq 0$, which is impossible since we have assumed that ϕ is a nonnegative function and is not identically zero. Therefore, there is no set $\mathcal{V}_F$ such that the representation formula (31) holds true for F.
3. The set $\mathcal{V}_F$ can be constructed as follows: Let C_p be defined as
$$\forall p \neq 0, \;\; C_p = \{w \in \mathcal{S}_n \mid \forall S \in \mathcal{S}_n, \; \mathrm{Tr}(Sw) \geq F(p, S)\},$$
and $C = \bigcup_{p\neq 0} C_p$. Then we can take
$$\mathcal{V}_F = \{v \in \mathcal{S}_n^+ \mid v^2 \in Co(C)\},$$
where $\mathcal{S}_n^+$ is the set of all $n \times n$ symmetric nonnegative matrices and where $Co(C)$ stands for the convex hull of the set C.

Proof. Let us first notice that, for any $p \neq 0$, the map $S \to F(p, S)$ is positively homogeneous. Indeed, from (27) and (30) (ii), we have

$$\forall S \in \mathcal{S}_n, \; \forall \lambda > 0, \; F(p, \lambda S) = \lambda F(p/\lambda, S) = \lambda F(p, S).$$

Let C_p be defined as in the remark. From the separation theorem applied to $F(p, \cdot)$, which is concave from assumption (30) (i) and positively homogeneous, the set C_p is nonempty, and we have

$$\forall S \in \mathcal{S}_n, \;\; F(p, S) = \inf_{w\in C_p} \mathrm{Tr}(Sw). \tag{33}$$

Let us also notice that, from assumption (30) (ii), we have

$$\forall p \neq 0,\ \forall \lambda \in \mathbb{R}\backslash\{0\},\ C_{\lambda p} = C_p. \tag{34}$$

Hence, from the definition of C_p, the set $\bigcup_{p\neq 0} C_p = \bigcup_{|p|=1} C_p$ is closed. Let us also point out that C_p is convex for any $p \neq 0$.

We claim that

$$\forall p \neq 0,\ \forall w \in C_p,\ \text{we have } w \geq 0 \text{ and} \quad \mathrm{Ker}(w) = \mathbb{R}p, \tag{35}$$

where Ker(w) stands for the kernel of w.

Proof of the claim. Let $w \in C_p$. For any $x \in \mathbb{R}^n$, we have, from (26) and (33) applied with $S = xx^*$,

$$0 = F(p, 0) \ \leq\ F(p, xx^*) \ \leq\ \langle wx, x\rangle.$$

Hence $w \geq 0$. Using assumption (27), (30) (ii), and (33), we get, for any $\lambda \in \mathbb{R}\backslash\{0\}$,

$$0 = F(p, \lambda pp^*) \ \leq\ \lambda\langle wp, p\rangle,$$

from which we deduce that $\mathbb{R}p \subset \mathrm{Ker}(w)$. Finally, from (30) (iii), we have, for $x \notin \mathbb{R}p$,

$$0 = F(p, 0) \ <\ F(p, xx^*) \ \leq\ \langle wx, x\rangle,$$

which implies that $\mathrm{Ker}(w) \subset \mathbb{R}p$.

We now claim that C_p is bounded independently of $p \neq 0$. Let $p \neq 0$ and $w \in C_p$. From (33) applied to $S = -I_n$, we have

$$\mathrm{Tr}(Sw) = -\mathrm{Tr}(w) \geq F(p, -I_n) \geq \min_{|q|=1} F(q, -I_n) > -\infty,$$

because F is continuous on $\mathbb{R}^n\backslash\{0\} \times \mathcal{S}_n$ (assumption (29)). From this inequality, we deduce that

$$\forall p \in \mathbb{R}^n\backslash\{0\},\ \forall w \in C_p, \quad \mathrm{Tr}(w) \leq -\min_{|q|=1} F(q, -I_n),$$

which in turn implies that C_p, for $p \neq 0$, are uniformly bounded since the w are nonnegative matrices. In particular, this proves that the set $C = \bigcup_{p\neq 0} C_p$ is compact.

We now claim that

$$\forall p \neq 0,\ \{w \in Co(C) \mid wp = 0\} = C_p. \tag{36}$$

Proof. Let us first notice that (35) implies that C_p is contained in $\{w \in Co(C) \mid wp = 0\}$. Let w belong to $Co(C)$ with $wp = 0$. There are $w_i \in C_{p_i}$ for some $p_i \neq 0$ and $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$, such that $w = \sum_i \lambda_i w_i$. Then $wp = 0$ implies that

$$0 = \left\langle \sum_i \lambda_i w_i p, p \right\rangle = \sum_i \lambda_i \langle w_i p, p\rangle.$$

From (35), $\langle w_i p, p\rangle \geq 0$ for any i, and thus $\langle w_i p, p\rangle = 0$. Then (35) again states that $p_i \in \mathbb{R}p$, which proves that, for any i, $w_i \in C_{p_i} = C_p$ from (34). Since C_p is convex, w belongs to C_p, and (36) is established.

Let us finally set

$$\mathcal{V}_F = \{v \in \mathcal{S}_n^+ \mid v^2 \in Co(C)\}.$$

Let us notice that

$$\{v^2 \mid v \in \mathcal{V}_F\} = Co(C)$$

because, for any $w \in Co(C)$, we have $w \geq 0$ from (35), and thus there is some $v \in \mathcal{S}_n^+$ with $v^2 = w$. Hence (32) is satisfied because $Co(C)$ is convex and compact. We now prove equality (31). We have

$$\forall (p, S) \in \mathbb{R}^n \backslash \{0\} \times \mathcal{S}_n, \quad \min_{v \in \mathcal{V}_F,\, vp=0} \mathrm{Tr}(Sv^2) = \min_{w \in Co(C),\, wp=0} \mathrm{Tr}(Sw) = \min_{w \in C_p} \mathrm{Tr}(Sw)$$

from (36). Thus we have, from (33),

$$\forall (p, S) \in \mathbb{R}^n \backslash \{0\} \times \mathcal{S}_n, \quad \min_{w \in C,\, vp=0} \mathrm{Tr}(Sv^2) = \min_{w \in C_p} \mathrm{Tr}(Sw) = F(p, S),$$

which proves (31). □

THEOREM 4.2. *Let us assume that F satisfies* (26), (27), (29), *and* (30). *Let $\mathcal{V}_F$ satisfy properties* (31) *and* (32) *of Lemma* 4.1, *and let us denote by $\mathcal{A}_F = \mathcal{A}_F(\Omega, \mathcal{F}, P; W)$ the set of all $\mathcal{V}_F$-valued $(\mathcal{F}_s)$-progressively measurable processes $v(\cdot)$. Then the following hold.*

1. *For any uniformly continuous, bounded function $g : \mathbb{R}^n \to \mathbb{R}^n$, the solution u of* (28) *can be represented as follows:*

$$\forall (t, x) \in \mathbb{R}_+ \times \mathbb{R}^n, \quad u(t, x) = \inf_{v(\cdot) \in \mathcal{A}_F} \text{ess-sup}_\Omega g(X^{x,v(\cdot)}(t)),$$

where $X^{x,v(\cdot)}$ is the solution to

$$\left\{ \begin{array}{l} dX^{x,v(\cdot)}(s) = \sqrt{2} v(s) dW(s), \\ X^{x,v(\cdot)}(0) = x. \end{array} \right. \tag{37}$$

2. *The front (Σ_t) associated with F can be represented as follows:*

$$\forall t \geq 0,\ \Sigma_t = \{x \in \mathbb{R}^n \mid \exists v(\cdot) \in \mathcal{A}_F \text{ such that } X^{x,v(\cdot)}(t) \in \Sigma_0 \text{ a.s.}\},$$

where $X^{x,v(\cdot)}$ is the solution to (37).

Remark. Distance solutions of the evolution equation associated with (25) can be characterized exactly as in Proposition 3.1.

Proof. It is the same proof as for Theorems 1.1 and 2.1. So we omit it. □

Appendix. A viability theorem. Let us consider a stochastic control system described by the following differential equation:

$$\left\{ \begin{array}{l} dX(s) = b(s, X(s), v(s))ds + \sigma(s, X(s), v(s))dW(s), \\ X(t) = x, \end{array} \right. \tag{38}$$

where $\mathcal{V}$ is a compact metric space, $b : [0, T] \times \mathbb{R}^n \times \mathcal{V} \to \mathbb{R}^n$, $\sigma : [0, T] \times \mathbb{R}^n \times \mathcal{V} \to \mathbb{R}^{n \times d}$, and W is now a d-dimensional standard Brownian motion. We denote by $X^{t,x,v(\cdot)}$ the solution of (38).

Let $(K_t)_{t \in [0,T]}$ be a family of moving sets. The aim of the following theorem is to give a necessary and sufficient condition for the existence of an admissible control $v(\cdot)$ that keeps $X^{t,x,v(\cdot)}(s)$ in K_s for $s \in [t, T]$, whenever $x \in K_t$.

In what follows, we assume that the control system satisfies the following conditions:

(H1) b and σ are uniformly continuous in (t,x,v);

(H2) $|\sigma(t,x,v)-\sigma(t,x',v)| \le C_0|x-x'|$ for all $t\in[0,T]$, for all $x,x'\in\mathbb{R}^n$, for all $v\in\mathcal{V}$;

(H3) $\langle b(t,x,v)-b(t,x',v),x-x'\rangle \le \mu|x-x'|^2$ for all $t\in[0,T]$, for all $x,x'\in\mathbb{R}^n$, for all $v\in\mathcal{V}$;

(H4) the set $\left\{\left(\frac{1}{2}\sigma\sigma^*(t,x,v),b(t,x,v)\right),\ v\in\mathcal{V}\right\}$ is convex and compact for all $t\in[0,T]$, for all $x\in\mathbb{R}^n$.

The following result is strongly inspired by several results already existing in the literature (see [1], [2], [4], [9]). We give here a version for time dependent sets.

THEOREM A.1. *Let $(K_t)_{t\in[0,T]}$ be a family of moving sets such that the set K defined by*

$$K=\bigcup_{t\in[0,T]}\{t\}\times K_t$$

is a closed subset of $[0,T]\times\mathbb{R}^n$. Let us assume that conditions (H1)$-$(H4) *are satisfied by the stochastic control system* (38)*. Then the following statements are equivalent:*

(i) *For any $t\in(0,T)$, for any $x\in K_t$, there is a control $v(\cdot)\in\mathcal{A}$ such that the solution $X^{t,x,v(\cdot)}$ to* (38) *satisfies*

$$\forall s\in[t,T],\ X^{t,x,v(\cdot)}(s)\in K_s \text{ a.s.}$$

(ii) *The map $u(t,x)=1-\mathbf{1}_{K_t}(x)$ is a supersolution to the following equation:*

$$u_t(t,x)+\inf_{v\in\mathcal{V},\ \sigma(t,x,v)^*Du(t,x)=0}\mathcal{L}_{t,x,v}u=0, \tag{39}$$

where $\mathcal{L}$ is the operator defined by

$$\mathcal{L}_{t,x,v}u=\langle b(t,x,v),Du(t,x)\rangle+\frac{1}{2}\mathrm{Tr}\left(D^2u\sigma\sigma^*\right)(t,x,v),$$

and where we use the convention $\inf_\emptyset=+\infty$.

(iii) *For any $\mathcal{C}^2$ function $\phi:[0,T]\times\mathbb{R}^n\to\mathbb{R}$ with a local maximum on K at (t,x), we have*

$$\phi_t(t,x)+\inf_{v\in\mathcal{V},\ \sigma(t,x,v)^*D\phi(t,x)=0}\left(\mathcal{L}_{t,x,v}\phi\right)\le 0.$$

Remark A.1.

1. The map $u(t,x)=1-\mathbf{1}_{K_t}(x)$ is lower semicontinuous (but not continuous in general) since the set K is closed.
2. Let us assume that the family (K_t) satisfies one of the above statements and the following regularity condition at time $t=0$:

$$\forall x\in K_0,\ \exists t_k\to 0,\ \exists x_k\in K_{t_k} \text{ with } x_k\to x. \tag{40}$$

Then (i) holds true up to time $t=0$.

3. The statements of Theorem A.1 are equivalent with the following assertion: For any $t\in(0,T)$, for any $x\in K_t$, there is a control $v(\cdot)\in\mathcal{A}$ such that the solution X of

$$\begin{cases} dX(s)=b(t+s,X(s),v(s))dt+\sigma(t+s,X(s),v(s))dW(s),\\ X(0)=x,\end{cases} \tag{41}$$

satisfies

$$\forall s\in[0,T-t],\ X(s)\in K_{t+s} \text{ a.s.}$$

Parts 2 and 3 of the remark are proved below.

Definition A.2. *If a family of moving sets* (K_t) *satisfies one of the equivalent conditions of Theorem* A.1, *we say that it is a viability domain for the controlled system* (38).

Proof of Theorem A.1. We are going to prove the following implications: (iii)⇒(ii) ⇒ (i)⇒(iii).

(iii)⇒(ii). This is straightforward, and we leave this proof to the reader.

(ii) ⇒ (i). Let us consider a uniformly continuous map $f : [0,T] \times \mathbb{R}^n \to \mathbb{R}$ such that

$$\forall (t,x) \in [0,T] \times \mathbb{R}^n,\ 0 \leq f(t,x) \leq 1, \quad \text{and} \quad f(t,x) = 0 \Leftrightarrow x \in K_t. \tag{42}$$

We introduce the following optimal control problem:

$$V(t,x) = \inf_{v \in \mathcal{A}} E \int_t^T f(s, X^{t,x,v(\cdot)}(s))ds.$$

It is known that $V(t,x)$ is the unique solution to the following equation:

$$V_t(t,x) + \inf_{v \in \mathcal{V}} (\mathcal{L}_{t,x,v} V)(t,x) + f(t,x) = 0 \tag{43}$$

with terminal condition $V(T,\cdot) = 0$ (cf. [8]). Moreover, the value function V is uniformly continuous and bounded. Finally, for any (t,x) there is at least one optimal control $v(\cdot)$, cf. [6]. Hence (i) holds true if and only if $V = 0$ on $K \cap (0,T)$.

Since u is a supersolution to (39), a straightforward computation shows that the map $w(t,x) = e^{(T-t)}u(t,x)$ satisfies (in the viscosity sense)

$$w_t + \inf_v \mathcal{L}_{t,x,v} w + w \leq 0.$$

Thus w is a supersolution of (43) because $f \leq w$ from (42). Moreover, w and V satisfy the following terminal conditions at time T:

$$w(T,\cdot) \geq 0 \quad \text{and} \quad V(T,\cdot) = 0.$$

Hence, from the comparison principle (see [8]), we have $w \geq V$ on $(0,T] \times \mathbb{R}^n$. Therefore, $V = 0$ on $K \cap (0,T)$, and implication (ii) ⇒ (i) is established.

(i) ⇒ (iii). Let ϕ be a test function with a local maximum on K at (t,x) with $t \in (0,T)$. This means that

$$\forall (s,y) \in K \cap B((t,x),r),\ \ \phi(s,y) \leq \phi(t,x) \tag{44}$$

for some $r > 0$, where $B((t,x),r) = [t-r, t+r] \times B(x,r)$ ($B(x,r)$ being the ball centered at x of radius r). From (i), there is some admissible control $v(\cdot) \in \mathcal{A}$ such that the solution $X^{t,x,v(\cdot)}$ satisfies

$$\forall s \in [t,T],\ \ X^{t,x,v(\cdot)}(s) \in K_s \quad \text{a.s.} \tag{45}$$

For any $\epsilon \in (0,r)$, let us introduce the stopping time τ_ϵ defined by

$$\tau_\epsilon = (t+\epsilon) \wedge \inf\{s \geq t\ ,\ |X^{t,x,v(\cdot)}(s) - x| > \epsilon\}.$$

Then, from (44) and (45), we have

$$\forall s > t,\ \ \phi(s \wedge \tau_\epsilon, X^{t,x,v(\cdot)}(s \wedge \tau_\epsilon)) \leq \phi(t,x).$$

From standard arguments, the fact that this inequality holds for any $s > t$ and any $\epsilon > 0$ implies that

$$\phi_t(t,x) + \inf_{v\in\mathcal{V}} (\mathcal{L}_{t,x,v}\phi)(t,x) \leq 0. \tag{46}$$

Let $\theta : \mathbb{R} \to \mathbb{R}$ be an increasing function such that $\theta'(\phi(t,x)) = 1$ and $\theta''(\phi(t,x)) = \alpha$, where $\alpha > 0$ is arbitrary. Since ϕ has a local maximum on K at (t,x), $\theta \circ \phi$ also has a local maximum on K at (t,x). Hence, from the definition of $\mathcal{L}$ and inequality (46), we have at the point (t,x)

$$\phi_t + \inf_{v\in\mathcal{V}} \left(\langle b, D\phi\rangle + \frac{1}{2}\mathrm{Tr}((\alpha D\phi D\phi^* + D^2\phi)\sigma\sigma^*) \right) \leq 0. \tag{47}$$

Let $v_\alpha \in \mathcal{V}$ be a minimum in the above expression. When $\alpha \to +\infty$, the expresssion

$$\mathrm{Tr}(\alpha D\phi(t,x)D\phi(t,x)^*\sigma(t,x,v_\alpha)\sigma^*(t,x,v_\alpha)) = \alpha|\sigma^*(t,x,v_\alpha)D\phi(t,x)|^2$$

remains bounded, which proves, since $\mathcal{V}$ is compact, that there are some $\alpha_k \to +\infty$ and some $v \in \mathcal{V}$ such that $v_{\alpha_k} \to v$ and $\sigma^*(t,x,v)D\phi(t,x) = 0$. Let us also point out that, for any k, inequality (47) implies that

$$\phi_t + \langle b(v_{\alpha_k}), D\phi\rangle + \frac{1}{2}\mathrm{Tr}(D^2\phi\sigma(v_{\alpha_k})\sigma^*(v_{\alpha_k})) \leq 0,$$

where we have omitted the dependence in (t,x) for simplicity. Letting $k \to +\infty$ gives

$$\phi_t + \langle b(v), D\phi\rangle + \frac{1}{2}\mathrm{Tr}(D^2\phi\sigma(v)\sigma^*(v)) \leq 0.$$

Since $\sigma^*(t,x,v)D\phi(t,x) = 0$, we have proved the desired result. □

Proof of Remark A.1. We first prove part 2. If (K_t) enjoys property (40), then the value function V, defined above in the proof of (ii)⇒(i), vanishes also on $\{0\} \times K_0$. Indeed, V is zero on $K \cap (0,T)$, and, from (40), any point of $\{0\} \times K_0$ can be approximated by a point of $K \cap (0,T)$. Hence (i) holds true up to time $t = 0$.

Let us now prove part 3 of the remark. We first assume that (K_t) is a viability domain for system (38). Let us fix some $t \in (0,T)$. Using characterization (ii) or (iii) of Theorem A.1, one can prove that the family $(\tilde{K}_s)_{s\in[0,T-t]}$ defined by

$$\forall s \in [0, T-t],\ \tilde{K}_s = K_{s+t}$$

is a viability domain for the stochastic control system

$$\begin{cases} d\tilde{X}(\tau) = \tilde{b}(\tau, \tilde{X}(\tau), v(\tau))d\tau + \tilde{\sigma}(\tau, \tilde{X}(\tau), v(\tau))dW(\tau), \\ \tilde{X}(s) = x, \end{cases} \tag{48}$$

where

$$\tilde{b}(\tau,y,v) = b(t+\tau,y,v) \quad \text{and} \quad \tilde{\sigma}(\tau,y,v) = \sigma(t+\tau,y,v).$$

We claim that $(\tilde{K}_s)_{s\in[0,T-t]}$ satisfies property (40). Indeed, for any $x \in \tilde{K}_0 = K_t$, there exists a control v for which the solution $X^{t,x,v}$ of (38) satisfies $X^{t,x,v}(s) \in K_s$ on $[t,T]$ a.s. Let us choose ω such that this property is satisfied and for which $s \to X^{t,x,v}_\omega(s)$ is continuous. Then, for any sequence $s_k \to t^+$, the sequence of points

$x_k = X_\omega^{t,x,v}(s_k)$ converges to x and belongs to K_{s_k}. Hence any point x of $\tilde{K}_0 = K_t$ can be approximated by a sequence of points x_k of $\tilde{K}_{s_k - t} = K_{s_k}$. This proves (40) for $\tilde{K}$. Now applying part 2 of Remark A.1 to the viability domain $(\tilde{K}_s)_{s\in[0,T-t]}$ gives, for any $x \in \tilde{K}_0$, the existence of some control $v(\cdot)$ such that the solution $\tilde{X}^{0,x,v(\cdot)}$ of (48) satisfies

$$\tilde{X}^{0,x,v(\cdot)}(\tau) \in \tilde{K}_\tau \quad \forall \tau \in [0, T-t] \text{ a.s.}$$

Since $\tilde{K}_\tau = K_{\tau+t}$ and since $\tilde{X}^{0,x,v(\cdot)}$ is nothing but the solution of (41), we have proved the desired result.

Let us now assume that the set (K_t) satisfies the conditions of part 3 of the remark. Let V be the value function defined in the above proof of (ii)⇒(i). Let us define V^t on $[0, T-t] \times \mathbb{R}^n$ by $V^t(s,y) = V(s+t, y)$. Then one easily checks that $V^t = \tilde{V}$, where $\tilde{V}$ is the value function of the following control problem:

$$\tilde{V}(s,y) = \inf_{v \in \mathcal{A}} E \int_s^{T-t} f(t+\tau, \tilde{X}^{s,y,v(\cdot)}(\tau))d\tau,$$

where $\tilde{X}^{s,y,v(\cdot)}$ is the solution to (48). Indeed, V^t and $\tilde{V}$ both satisfy the same equation on $(0, T-t) \times \mathbb{R}^n$ (namely, (43) with $\tilde{f}(s,\cdot) = f(s+t,\cdot)$ instead of f) and have the same terminal condition $V^t(T-t,\cdot) = \tilde{V}(T-t,\cdot) = 0$. Moreover, our assumption implies that $V^t = \tilde{V}$ vanishes on $\{0\} \times \tilde{K}_0$. This means that the value function V vanishes on $\{t\} \times K_t$. This in turn implies that, for any $x \in K_t$, there is an optimal control v for $V(t,x)$ (cf. [6]) such that the solution $X^{t,x,v}$ of (38) satisfies $X^{t,x,v}(s) \in K_s$ on $[t,T]$ a.s. Since this holds true for any $t \in (0,T)$, we have proved that (K_t) is a viability domain for (38). □

Acknowledgment. We would like to thank J.-P. Aubin for suggesting the characterization in terms of viability given in section 3.

REFERENCES

[1] J.-P. Aubin and G. Da Prato, *Stochastic viability and invariance*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 17 (1990), pp. 595–613.

[2] J.-P. Aubin and G. Da Prato, *The viability theorem for stochastic differential inclusions*, Stochastic Anal. Appl., 16 (1997), pp. 1–15.

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

[4] R. Buckdahn, S. Peng, M. Quincampoix, and C. Rainer, *Existence of stochastic control under state constraints*, C. R. Acad. Sci. Paris Sér. I Math., 327 (1998), pp. 17–22.

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

[6] N. El Karoui, D. Hu Nguyen, and M. Jeanblanc-Piqué, *Compactification methods in the control of degenerate diffusions: Existence of an optimal control*, Stochastics, 20 (1987), pp. 169–219.

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

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

[9] S. Gautier and L. Thibault, *Viability for constrained stochastic differential equations*, Differential Integral Equations, 6 (1993) pp. 1395–1414.

[10] Y. Giga, S. Goto, H. Ishii, and M.-H. Sato, *Comparison principle and convexity preserving properties for singular degenerate parabolic equations on unbounded domains*, Indiana Univ. Math. J., 40 (1990), pp. 443–470.

[11] H. M. Soner, *Motion of a set by the mean curvature of its boundary*, J. Differential Equations, 101 (1993), pp. 313–372.

[12] H. M. Soner and N. Touzi, *Dynamic Programming for Stochastic Target Problems and Geometric Flows*, preprint, CERMSEM, Université Paris I, Paris, France, 2000.

[13] J. Yong and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York, 1999.

SIAM J. Math. Anal.
Vol. 33, No. 4, pp. 847–859

MATHEMATICAL JUSTIFICATION OF THE HYDROSTATIC APPROXIMATION IN THE PRIMITIVE EQUATIONS OF GEOPHYSICAL FLUID DYNAMICS*

PASCAL AZÉRAD† AND FRANCISCO GUILLÉN‡

Abstract. Geophysical fluids all exhibit a common feature: their aspect ratio (depth to horizontal width) is very small. This leads to an asymptotic model widely used in meteorology, oceanography, and limnology, namely the hydrostatic approximation of the time-dependent incompressible Navier–Stokes equations. It relies on the hypothesis that pressure increases linearly in the vertical direction. In the following, we prove a convergence and existence theorem for this model by means of anisotropic estimates and a new time-compactness criterium.

Key words. Navier–Stokes equations, shallow domains, geophysical fluid dynamics, hydrostatic approximation, singular perturbation, compactness criterium, asymptotic analysis

AMS subject classifications. 35Q30, 35B40, 76D05, 34C35

PII. S0036141000375962

1. Introduction.

Atmospheric flow in meteorology, water flow in oceanography, and limnology are all described by the Navier–Stokes equations. Due to the fact that the aspect ratio

$$\epsilon = \frac{\text{characteristic depth}}{\text{characteristic width}}$$

is very small in most geophysical domains, asymptotic models have been used; see, e.g., [9, 15, 22]. One such model is the primitive equations model; see, e.g., [11, 12], wherein the unknown flow variables are velocity, pressure, temperature, and salinity (in the case of an ocean). Besides, most geophysical fluids are stratified (i.e., density is a known function of the temperature (and salinity, if any)) and have a free surface. We shall not investigate these features in this paper, leaving it, rather, for forthcoming work.

Instead we shall focus on the assumption that the pressure is hydrostatic, i.e., increases linearly with respect to the depth, as in the static case. This law agrees well with experiment (as first observed by Blaise Pascal around 1650; see [14])) and is frequently taken as a hypothesis in geophysical fluid dynamics. We justify this assumption by means of asymptotic analysis (taking ϵ as the small parameter). Our derivation is made possible by the use of anisotropic eddy viscosities, namely $\nu = (\nu_x, \nu_y, \nu_z)$, relying on the fact that the ratio between the horizontal and vertical scales leads to very different sizes for the horizontal and vertical eddies (see [9, 15]). Specifically, if we assume that $\nu = (\nu_1, \nu_2, \epsilon^2 \nu_3)$ with $\nu_i = \mathrm{O}(1)$ for $i = 1, 2, 3$, then we will see that weak solutions of the Navier–Stokes equations converge to a weak solution of a limit problem with hydrostatic pressure.

*Received by the editors July 26, 2000; accepted for publication (in revised form) August 22, 2001; published electronically December 18, 2001. This work was supported by the fonds Franco-Espagnol D.R.E.I.F. (ref. UC 815) and the project MAR98-0486 C.I.C.Y.T. (España).

http://www.siam.org/journals/sima/33-4/37596.html

†Laboratoire de Modélisation, Analyse Non Linéaire et Optimisation, Université de Perpignan, 52 av. de Villeneuve, F-66860 Perpignan cedex, France (azerad@univ-perp.fr).

‡Departamento de Ecuaciones Diferenciales y Análisis Numérico, c/ tarfia s/n Universidad de Sevilla, 41012 Sevilla, Spain (guillen@numer.us.es).

The stationary case has already been studied (see [4] for the linear problem and [5] for the nonlinear one), whereas the linear time-dependent case was solved in [1]. The main task of this paper is then to solve the nonlinear time-dependent case. Our result was announced in [2], whereas numerical simulations stemming from it were discussed in [3].

Fluid flow in thin domains (flat, curved, and with various boundary conditions) has been extensively studied; see [7, 13, 16, 20, 21]. In these works, an isotropic viscosity is used, and the depth is constant. By averaging along the vertical direction, two-dimensional (2D) limit models are obtained, together with existence and global regularity results.

Our approach is different, because we neither eliminate the vertical velocity by averaging nor assume the depth of the domain to be constant. By making use of different horizontal and vertical eddy viscosities, we are able to derive a three-dimensional (3D) limit nonlinear model. Let us emphasize that the anisotropic viscosity hypothesis is fundamental for the derivation of the primitive equations: in the stationary case, keeping an isotropic viscosity, the asymptotic model is linear, with vanishing horizontal diffusion; see [6].

The paper is organized as follows. In section 2, we present the physical model and the scaling leading to the primitive equations. We state the main theorem in section 3. The functional setting and weak formulation are described in section 4. In the next section, we state and prove a time-compactness result, which we shall use in the proof of the main theorem in section 6. Finally, in section 7, we comment on the convergence of the pressure and the orders of magnitude of the vertical velocity with respect to the aspect ratio.

2. Equations governing the flow and scaling. Let us consider an incompressible homogeneous fluid filling a thin domain defined by

$$\Omega_\epsilon = \left\{(x,y,z) \in \mathbb{R}^3;\ (x,y) \in \omega, -\epsilon\, h(x,y) < z < 0\right\},$$

where ω is an open bounded Lipschitz domain in $\mathbb{R}^2$ and $h : \overline{\omega} \to \mathbb{R}$ is a nonnegative lipschitzian application, which is arbitrary provided that Ω_ϵ is lipschitzian. In particular, h may vanish, contrary to [12, 9], but in order that the domain Ω_ϵ has no cusps, the slope must not vanish on the shores.[1] We denote by $\Gamma_s = \overline{\omega} \times \{0\}$ the fluid surface and by $\Gamma_b^\epsilon = \partial\Omega_\epsilon \setminus \Gamma_s$ the basin bottom. The fluid flow in Ω_ϵ is generated by the wind traction on the surface Γ_s, influenced by the Coriolis and centrifugal forces and governed by the Navier–Stokes equations, in which we take different eddy viscosities according to the direction; see [5, 9, 15]. Finally, we take the density as identically equal to one. In a geophysical rotating frame (z pointing upwards, x east, and y north), the initial-boundary value problem reads as follows.

Find $\mathbf{v} = (v_1, v_2, v_3)$ (velocity) and q (pressure), such that

$$\partial_t \mathbf{v} + (\mathbf{v}\cdot\nabla)\mathbf{v} - \Delta_\nu \mathbf{v} + \nabla q + 2\mathbf{w} \times \mathbf{v} = \mathbf{g} \quad \text{in } \Omega_\epsilon \times (0, \mathrm{T}), \tag{2.1}$$

$$\operatorname{div} \mathbf{v} = 0 \quad \text{in } \Omega_\epsilon \times (0, T), \tag{2.2}$$

$$\mathbf{v} = 0 \quad \text{on } \Gamma_b^\epsilon \times (0, T), \tag{2.3}$$

$$\nu_z \partial_z v_1 = \tau_1, \quad \nu_z \partial_z v_2 = \tau_2, \quad v_3 = 0 \quad \text{on } \Gamma_s \times (0, T), \tag{2.4}$$

$$\mathbf{v}(\cdot, t = 0) = \mathbf{v}_0 \quad \text{in } \Omega_\epsilon. \tag{2.5}$$

[1] This is a technical hypothesis. One could probably dispense with it due to the specific shape of the domain.

In (2.1), $\nabla = (\partial_x, \partial_y, \partial_z)$ denotes the gradient vector, and Δ_ν denotes the anisotropic Laplacian defined by $\Delta_\nu = \nu_x \partial^2_{xx} + \nu_y \partial^2_{yy} + \nu_z \partial^2_{zz}$ with $\nu = (\nu_x, \nu_y, \nu_z)$ the eddy kinematic viscosity vector. Moreover, $\mathbf{w} = f\,(0, \cos(l(y)), \sin(l(y)))$ represents the earth rotation angular speed (f the module and $l(y)$ the latitude), $2\mathbf{w} \times \mathbf{v}$ represents the Coriolis acceleration ($\times$ denotes the cross-product in $\mathbb{R}^3$), and $\mathbf{g}$ represents the force due to gravity (which also includes the centrifugal effect). It is well known (cf. [15, p. 18]) that $\mathbf{g}$ is a potential, i.e., $\mathbf{g} = \nabla\varphi$. It is customary to incorporate the gravity potential in the pressure term; thus we set

$$p = q - \varphi.$$

Equation (2.2) represents the incompressibility condition, and (2.3) represents the no-slip condition on the bottom.

In (2.4), τ_i, $i = 1, 2$, stand for the horizontal tractions exerted by the wind on the (fixed) surface Γ_s of the fluid, and $w = 0$ on Γ_s comes from the rigid lid hypothesis. In (2.5), $\mathbf{v}_0 = (v_{01}, v_{02}, v_{03})$ designates the initial velocity.

Remark. We have neglected the earth's curvature, and hence our analysis is valid only locally, e.g., for lakes; for seas or oceans, spherical coordinates should be used [12], although this can be somewhat cumbersome.

As usual in asymptotic analysis, we perform a vertical scaling to make the domain independent of ϵ, that is,

$$x = x_1, \quad y = x_2, \quad z = \epsilon\, x_3,$$

so that $\Omega = \big\{(x_1, x_2, x_3) \in \mathbb{R}^3;\ (x_1, x_2) \in \omega,\ -h(x_1, x_2) < x_3 < 0\big\}$ is the new fixed domain.

The corresponding kinematic scaling is

$$v_1 = u_1^\epsilon, \quad v_2 = u_2^\epsilon, \quad v_3 = \epsilon\, u_3^\epsilon, \quad p = p^\epsilon, \tag{2.6}$$

so that $\mathbf{u}^\epsilon = (u_1^\epsilon, u_2^\epsilon, u_3^\epsilon)$ is the new unknown velocity and p^ϵ is the new pressure.

It is necessary to scale the mechanical quantities accordingly. First, it is only natural to assume $v_{01} = u_{01}, v_{02} = u_{02}$, and $v_{03} = \epsilon u_{03}$, where u_{0i} does not depend on ϵ, $i = 1, 2, 3$. Next we assume $\nu_x = \nu_1$, $\nu_y = \nu_2$, and $\nu_z = \epsilon^2 \cdot \nu_3$, where ν_1, ν_2, ν_3 are constants. As mentioned in the introduction, in oceanography the vertical eddy viscosity is usually very small compared to the horizontal one. We refer to [5] for a mathematical discussion of this assumption, and here we content ourselves with one heuristic comment. Basically, a kinematic viscosity has the dimension L^2/T, where L (resp., T) is a typical length (resp., time) scale so that ν_x and ν_y have the dimension L_H^2/T, whereas ν_z has the dimension L_V^2/T, where L_H (resp., L_V) denotes a typical horizontal (resp., vertical) length scale. It follows that the ratio ν_z/ν_x and $\nu_z/\nu_y = \mathrm{O}(\epsilon^2)$.[2]

Now (2.4) becomes

$$\nu_3 \partial_3 u_i^\epsilon = \tau_i^\epsilon/\epsilon, \quad i = 1, 2.$$

We see that in order to end up with an O(1)-wind force on the rescaled domain, we have to assume that $\tau_i^\epsilon = \epsilon \cdot \theta_i$, $i = 1, 2$, where the θ_i are functions independent of ϵ.

[2] We do not delude ourselves with this sketchy argument. As far as we know, up to now there has been no rigorous derivation of any eddy viscosity model.

Remark. This last assumption can also be motivated by dimensional analysis, as follows. From $\tau_i = \nu_z \partial_z v_i$, one derives that τ_i has the dimension of

$$\frac{L_V^2}{T} \cdot \frac{1}{L_V} \cdot \frac{L_H}{T} = \epsilon \frac{L_H^2}{T^2} = O(\epsilon).$$

With the above considerations, problem (2.1)–(2.5) transforms into the following anisotropic Navier–Stokes equations:

$$\partial_t u_1^\epsilon + \mathbf{u}^\epsilon \cdot \nabla u_1^\epsilon - \Delta_\nu u_1^\epsilon - \alpha\, u_2^\epsilon + \epsilon\, \beta\, u_3^\epsilon + \partial_1 p^\epsilon = 0 \quad \text{in } \Omega \times (0,T), \tag{2.7}$$

$$\partial_t u_2^\epsilon + \mathbf{u}^\epsilon \cdot \nabla u_2^\epsilon - \Delta_\nu u_2^\epsilon + \alpha\, u_1^\epsilon + \partial_2 p^\epsilon = 0 \quad \text{in } \Omega \times (0,T), \tag{2.8}$$

$$\epsilon^2 \left\{ \partial_t u_3^\epsilon + \mathbf{u}^\epsilon \cdot \nabla u_3^\epsilon - \Delta_\nu u_3^\epsilon \right\} - \epsilon\, \beta\, u_1^\epsilon + \partial_3 p^\epsilon = 0 \quad \text{in } \Omega \times (0,T), \tag{2.9}$$

$$\operatorname{div} \mathbf{u}^\epsilon = 0 \quad \text{in } \Omega \times (0,T), \tag{2.10}$$

$$\mathbf{u}^\epsilon = 0 \quad \text{on } \Gamma_b \times (0,T), \tag{2.11}$$

$$\nu_3 \partial_3 u_1^\epsilon = \theta_1, \quad \nu_3 \partial_3 u_2^\epsilon = \theta_2, \quad u_3^\epsilon = 0 \quad \text{on } \Gamma_s \times (0,T), \tag{2.12}$$

$$\mathbf{u}^\epsilon(\cdot, t=0) = \mathbf{u}_0 \quad \text{in } \Omega. \tag{2.13}$$

Now $\nabla = (\partial_1, \partial_2, \partial_3)$, $\Delta_\nu = \nu_1 \partial_{11}^2 + \nu_2 \partial_{22}^2 + \nu_3 \partial_{33}^2$, $\Gamma_b = \partial\Omega \setminus \Gamma_s$, $\alpha = 2f \sin(l(x_2))$, and $\beta = 2f \cos(l(x_2))$.

If we assume that $\mathbf{u}^\epsilon = O(1)$, then neglecting the ϵ^2 and ϵ terms in the first and third momentum equation, (2.7) and (2.9), we formally get the hydrostatic Navier–Stokes equations, also called the primitive equations:

$$\partial_t u_1 + \mathbf{u} \cdot \nabla u_1 - \Delta_\nu u_1 - \alpha\, u_2 + \partial_1 p = 0 \quad \text{in } \Omega \times (0, \mathrm{T}), \tag{2.14}$$

$$\partial_t u_2 + \mathbf{u} \cdot \nabla u_2 - \Delta_\nu u_2 + \alpha\, u_1 + \partial_2 p = 0 \quad \text{in } \Omega \times (0, \mathrm{T}), \tag{2.15}$$

$$\partial_3 p = 0 \quad \text{in } \Omega \times (0, \mathrm{T}), \tag{2.16}$$

$$\operatorname{div} \mathbf{u} = 0 \quad \text{in } \Omega \times (0,T), \tag{2.17}$$

$$u_1 \, = \, u_2 \, = u_3 n_3 = 0 \quad \text{on } \Gamma_b \times (0,T), \tag{2.18}$$

$$\nu_3 \partial u_1 = \theta_1, \quad \nu_3 \partial_3 u_2 = \theta_2, \quad u_3 = 0 \quad \text{on } \Gamma_s \times (0,T), \tag{2.19}$$

$$u_i(\cdot, t=0) = u_{0i} \quad \text{in } \Omega,\ i = 1, 2. \tag{2.20}$$

Remark. The boundary condition (2.18) differs from its counterpart (2.11) because u_3 is less regular than u_1, u_2 as we shall see below. Also, the initial condition (2.20) does not involve u_3, the time derivative of which is missing in the hydrostatic model. The problem is not in the Cauchy–Kowalevska form.[3]

Remark. If u_3 were to be computed directly from (2.17), which is a *first* order equation, it is not obvious at all that it would fulfill the *two* boundary conditions on the bottom (2.18) and the surface (2.19).

3. Main theorem. Let T be a fixed positive duration. We make the natural assumption of a wind of finite energy: θ_1, $\theta_2 \in L^2(0,T; H^{-1/2}(\Gamma_s))$. Our main result is the following theorem.

THEOREM 3.1. *Let* $\mathbf{u}_0 \in L^2(\Omega)^3$, *with* $\operatorname{div} \mathbf{u}_0 = 0$, $\mathbf{u}_0 \cdot \mathbf{n} = 0$ *on* $\partial\Omega$, *and* $\theta_1, \theta_2 \in L^2(0,T; H^{-1/2}(\Gamma_s))$*; there exists a weak solution* $\mathbf{u}$ *of the hydrostatic Navier–Stokes equations* (2.14)–(2.20)*, obtained as a limit of weak solutions* $\mathbf{u}^\epsilon$ *of the anisotropic Navier–Stokes equations* (2.7)–(2.13)*, as the aspect ratio* ϵ *tends to zero.*

[3] Meteorologists say that u_3 is no longer a prognostic variable (see [11, 12]).

The proof relies on a priori estimates in anisotropic spaces (Propositions 6.1 and 6.2), which are sufficient to take the limit in the linear terms (see [1]), whereas for the nonlinear terms, we establish a new time-compactness criterium (Theorem 5.1), which enables us to get strong convergence of the horizontal velocities; see Lemma 6.3. This theorem states essentially that a small perturbation of an L^p-equicontinuous family still possesses a strong convergent subsequence. Let us emphasize that this seemingly technical refinement is by no means superfluous. Indeed, the usual compactness estimate fails: as $(u_1^\epsilon, u_2^\epsilon, \epsilon^2 u_3^\epsilon)$ is not divergence free, even if it is easy from (2.7)–(2.9) to control $\partial_t(u_1^\epsilon, u_2^\epsilon, \epsilon^2 u_3^\epsilon)$ in some dual space of divergence free velocities, it is not possible to apply the Aubin–Lions lemma to get compactness.

Another major difficulty of the proof is the lack of regularity of the vertical velocity, which is determined only by the incompressibility equation (2.10).

Remark. It is possible to handle a general force (f_1, f_2, f_3) in problem (2.14)–(2.20), by simply adding $\mathbf{f} = (f_1, f_2, \frac{f_3}{\epsilon})$ to (2.1), in order to end up with (f_1, f_2, f_3) in (2.7)–(2.9).

4. Weak formulation and anisotropic spaces. We need the following Hilbert spaces:

$$H_b^1(\Omega) = \overline{C_b^\infty(\Omega)}^{H^1(\Omega)} = \left\{v \in H^1(\Omega);\ v = 0 \text{ on } \Gamma_b\right\}$$

(where $C_b^\infty(\Omega) = \left\{\varphi \in C^\infty(\bar{\Omega});\ \varphi = 0 \text{ in some neighborhood of } \Gamma_b\right\}$),

$$\mathbf{V} = \left\{\mathbf{v} \in H_b^1(\Omega) \times H_b^1(\Omega) \times H_0^1(\Omega);\ \operatorname{div} \mathbf{v} = 0 \text{ in } \Omega\right\},$$

$$H(\partial_3, \Omega) = \left\{v \in L^2(\Omega);\ \partial_3 v \in L^2(\Omega)\right\}$$

(endowed with the norm $\|v\|^2_{H(\partial_3,\Omega)} = \|v\|^2_{L^2(\Omega)} + \|\partial_3 v\|^2_{L^2(\Omega)}$),

$$H_0(\partial_3, \Omega) = \overline{C_0^\infty(\Omega)}^{H(\partial_3,\Omega)} = \{v \in H(\partial_3, \Omega);\ v\, n_3 = 0 \text{ on } \partial\Omega\}$$

(n_3 is the third component of the normal exterior vector on $\partial\Omega$, and $v\, n_3$ is understood in the $H^{-1/2}(\partial\Omega)$ sense (see [19] for these spaces)),

$$\mathbf{W} = \left\{\mathbf{u} \in H_b^1(\Omega) \times H_b^1(\Omega) \times H_0(\partial_3, \Omega);\ \operatorname{div} \mathbf{u} = 0 \text{ in } \Omega\right\}.$$

Let us denote that $u_H = (u_1, u_2)$, $\theta_H = (\theta_1, \theta_2)$, $b(u_H) = \alpha\,(-u_2, u_1)$, and $\nabla_\nu = (\nu_1^{1/2}\partial_1, \nu_2^{1/2}\partial_2, \nu_3^{1/2}\partial_3)$. The scalar product in $L^2(\Omega)^d$, or the duality $L^p(\Omega), L^{p'}(\Omega)$, is denoted by $(\cdot,\cdot)$, and the duality $H^{-1/2}(\Gamma_s)H^{1/2}(\Gamma_s)$, is denoted by $\langle\cdot,\cdot\rangle_{\Gamma_s}$.

The weak form of the hydrostatic Navier–Stokes equations (2.14)–(2.20) is then as follows.

Find $\mathbf{u} = (u_H, u_3) \in L^2(0, T; \mathbf{W})$, with $u_H \in L^\infty(0, T; L^2(\Omega)^2)$, such that

$$\begin{aligned} \int_0^T -(u_H, \partial_t v_H) - (u_H, (\mathbf{u}\cdot\nabla)v_H) + (b(u_H), v_H) + (\nabla_\nu u_H, \nabla_\nu v_H) \\ = -(u_{0H}, v_H(0)) + \int_0^T \langle\theta_H, v_H\rangle_{\Gamma_s} \end{aligned} \tag{4.1}$$

for all $\mathbf{v} = (v_H, v_3) \in H^1(0, T; \mathbf{W})$, with $v_H(T) = 0$ and $\partial_3 v_H \in L^\infty(0, T; L^3(\Omega)^2)$.

Remark. Notice that a weak solution of the Navier–Stokes equations verifies the following regularity:

$$\mathbf{u} \in L^2(0,T;\mathbf{V}) \cap L^\infty(0,T;L^2(\Omega)^3)$$

(cf. [8, 10, 18]). Now the lack of regularity of u_3 makes it necessary to change $\mathbf{V}$ to $\mathbf{W}$. Moreover, in general, $u_3 \notin L^\infty(0,T;L^2(\Omega))$.

Remark. The regularity $L^\infty(0,T;L^3(\Omega)^2)$ is required for $\partial_3 v_H$ to give a meaning to $\int_0^T (u_H, u_3\partial_3 v_H)\,\mathrm{dt}$. The regularity $L^2(0,T;L^\infty(\Omega)^2)$ or any interpolated one $L^{2/a}(0,T;L^{3/(1-a)}(\Omega)^2)$ with $0 \le a \le 1$ can also be considered.

5. Compactness by perturbation. We give a compactness criterium, new to our knowledge, which generalizes the well-known translation criterium of Riesz–Fréchet–Kolmogorov, extended to the vectorial case by Simon [17]. In the following, $\tau_h f(t)$ denotes $f(t+h)$.

THEOREM 5.1. *Let $T > 0$, and let the Banach spaces $\mathbf{X} \overset{compact}{\hookrightarrow} \mathbf{B} \hookrightarrow \mathbf{Y}$. Let $(f_\epsilon)_{\epsilon>0}$ be a family of functions of $L^p(0,T;\mathbf{X})$, $1 \le p \le \infty$, with the extra condition $(f_\epsilon)_{\epsilon>0} \subset \mathcal{C}(0,T;\mathbf{Y})$ if $p = \infty$, such that*
(H1) *$(f_\epsilon)_{\epsilon>0}$ is bounded in $L^p(0,T;\mathbf{X})$,*
(H2) *$\|\tau_h f_\epsilon - f_\epsilon\|_{L^p(0,T-h;\mathbf{Y})} \le \varphi(h) + \psi(\epsilon)$ with*

$$\begin{cases} \lim_{h\to 0}\varphi(h) = 0, \\ \lim_{\epsilon\to 0}\psi(\epsilon) = 0. \end{cases}$$

Then the family $(f_\epsilon)_{\epsilon>0}$ possesses a cluster point in $L^p(0,T;\mathbf{B})$ and also in $\mathcal{C}(0,T;\mathbf{B})$ if $p = \infty$, as $\epsilon \to 0$.

Proof. It is enough to prove that, for every sequence $(\epsilon_n)_n$ such as $\epsilon_n > 0$ and $\epsilon_n \to 0$, the family $(f_{\epsilon_n})_n$ is relatively compact in $L^p(0,T;\mathbf{B})$. We apply Theorem 5 of Simon [17, p. 84] to the sequence $(f_{\epsilon_n})_n$, while observing that hypothesis (H2) implies that

$$\|\tau_h f_{\epsilon_n} - f_{\epsilon_n}\|_{L^p(0,T-h;\mathbf{Y})} \to 0 \quad \text{as } h \to 0$$

uniformly with respect to n. Indeed, (H2) implies that

$$\forall n,\ \|\tau_h f_{\epsilon_n} - f_{\epsilon_n}\|_{L^p(0,T-h;\mathbf{Y})} \le \varphi(h) + \psi(\epsilon_n).$$

Let $\epsilon > 0$ and then $\exists\, N$, such that for all $n \ge N$, $\psi(\epsilon_n) \le \epsilon/2$. On the other hand, $\exists\, \delta > 0$, such that for all $h : 0 \le h < \delta$, $\varphi(h) \le \epsilon/2$. Therefore, we get the estimate

$$\forall\, n \ge N \text{ and } \forall\, h : 0 \le h < \delta, \quad \|\tau_h f_{\epsilon_n} - f_{\epsilon_n}\|_{L^p(0,T-h;\mathbf{Y})} \le \epsilon.$$

In addition, for each $k \le N$, $\exists\, \delta_k > 0$, such that for all $h : 0 \le h < \delta_k$

$$\|\tau_h f_{\epsilon_k} - f_{\epsilon_k}\|_{L^p(0,T-h;\mathbf{Y})} \le \epsilon.$$

This follows from the L^p-continuity by translation of an L^p function for $p < \infty$ and for $p = \infty$; this is precisely a hypothesis.

Defining $\eta = \min\{\delta, \delta_1, \dots, \delta_N\}$, we obtain the desired uniform estimate

$$\forall\, h : 0 \le h < \eta, \quad \|\tau_h f_{\epsilon_n} - f_{\epsilon_n}\|_{L^p(0,T-h;\mathbf{Y})} \le \epsilon \quad \forall n.$$

The family $(f_{\epsilon_n})_n$ fulfills the hypotheses of Simon's theorem. □

6. Proof of the main theorem.

For simplicity in the notation, from now on, unless we specify otherwise, we will denote $\mathbf{u} = \mathbf{u}^\epsilon$ as a weak solution of the anisotropic Navier–Stokes equations (2.7)–(2.13).

6.1. Energy estimates.

The usual energy inequality (cf. [10]) for the Navier–Stokes equations gives, for a.e. $t \in [0, T]$,

$$\begin{aligned} &\|u_H(t)\|_{L^2}^2 + \epsilon^2 \|u_3(t)\|_{L^2}^2 + \int_0^t \{\|\nabla_\nu u_H(\tau)\|_{L^2}^2 + \epsilon^2 \|\nabla_\nu u_3(\tau)\|_{L^2}^2\} \, d\tau \\ &\quad \le \|u_{0H}\|_{L^2}^2 + \epsilon^2 \|u_{03}\|_{L^2}^2 + \int_0^t \langle \theta_H, u_H \rangle_{\Gamma_s}. \end{aligned}$$

Hence we obtain as in the isotropic Navier–Stokes system (cf. [1]) the following proposition.

Proposition 6.1. *The sequences $u_1, u_2, \epsilon u_3$ are bounded in $L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega))$.*

For the vertical velocities, we prove the following.

Proposition 6.2. *The sequences u_3 and $\partial_3 u_3$ are bounded in $L^2(0, T; L^2(\Omega))$; i.e., u_3 is bounded in $L^2(0, T; H_0(\partial_3, \Omega))$.*

Proof. As $\operatorname{div} \mathbf{u} = 0$, $\partial_3 u_3 = -\partial_1 u_1 - \partial_2 u_2$ is bounded in $L^2(0, T; L^2(\Omega))$. Moreover, the Poincaré inequality in the vertical direction, owing to $u_3 = 0$ on Γ_s, yields

$$\|u_3\|_{L^2} \le h_{\max} \|\partial_3 u_3\|_{L^2}, \quad \text{where } h_{\max} = \max_{\overline{\omega}} h.$$

Therefore, we have proved the proposition. □

6.2. Fractional time derivatives in horizontal spaces.

First, we define the auxiliary Hilbert spaces

$$B_H = \overline{P_H \mathcal{U}}^{(L^2)^2}, \quad W_H = \overline{P_H \mathcal{U}}^{(H^1)^2}, \quad \text{and} \quad Y_H = \overline{P_H \mathcal{U}}^{(H^2)^2},$$

where

$$\mathcal{U} = \left\{\varphi \in C_b^\infty(\Omega)^2 \times C_0^\infty(\Omega); \ \operatorname{div} \varphi = 0\right\}$$

and P_H is the projection

$$P_H : (x_1, x_2, x_3) \in \mathbb{R}^3 \mapsto (x_1, x_2) \in \mathbb{R}^2.$$

Then, from the Sobolev–Rellich embeddings, one deduces easily that

$$Y_H \hookrightarrow W_H \hookrightarrow B_H \equiv B_H' \hookrightarrow W_H' \hookrightarrow Y_H', \tag{6.1}$$

where all are dense and compact embeddings. Here and henceforth, X' denotes the dual space of X.

Now, we have the following lemma.

Lemma 6.3. *The estimate $\|\tau_h u_H - u_H\|_{L^\infty(0, T-h; Y_H')} \le C(h^{1/4} + \epsilon)$ holds.*

Proof. The spatial weak form of the Navier–Stokes equation (2.7)–(2.13) is

$$\begin{aligned} &\frac{d}{dt}(u_H, v_H) - (u_H, (\mathbf{u} \cdot \nabla) v_H) + (b(u_H), v_H) + (\nabla_\nu u_H, \nabla_\nu v_H) \\ &\quad + \epsilon^2 \left\{ \frac{d}{dt}(u_3, v_3) + (\mathbf{u} \cdot \nabla u_3, v_3) + (\nabla_\nu u_3, \nabla_\nu v_3) \right\} \\ &\quad + \epsilon \left\{ (\beta u_3, v_1) - (\beta u_1, v_3) \right\} = \langle \theta_H, v_H \rangle_{\Gamma_s} \quad \text{in } \mathcal{D}'(0, T) \\ &\quad \forall \mathbf{v} = (v_H, v_3) \in \mathbf{V}. \end{aligned} \tag{6.2}$$

Letting $v_H \in Y_H$, there is a null divergence lifting $\mathbf{v} = (v_H, v_3) \in H^2_b(\Omega)^2 \times H^1_0(\partial_3, \Omega)$ such that

$$\|v_3\|_{H^1} + \|\partial_3\, v_3\|_{H^1} \le C\|v_H\|_{Y_H}. \tag{6.3}$$

Here, the spaces $H^2_b(\Omega)$ and $H^1_0(\partial_3, \Omega)$ are the natural extensions of the spaces $H^1_b(\Omega)$ and $H_0(\partial_3, \Omega)$:

$$H^2_b(\Omega) = \overline{C^\infty_b(\Omega)}^{H^2(\Omega)} = \left\{ v \in H^2(\Omega);\ v = \frac{\partial v}{\partial n} = 0 \text{ on } \Gamma_b \right\},$$

$$H^1(\partial_3, \Omega) = \left\{ v \in H^1(\Omega);\ \partial_3 v \in H^1 \Omega) \right\},$$

$$H^1_0(\partial_3, \Omega) = \overline{C^\infty_0(\Omega)}^{H^1(\partial_3,\Omega)} = \left\{ v \in H^1(\partial_3, \Omega);\ v = \partial_3 v = \ 0 \ \text{ on } \partial\Omega \right\}.$$

Indeed, as $v_H \in Y_H$, there exists a sequence $(\varphi^n_H, \varphi^n_3) \in \mathcal{U}$ such that $\varphi^n_H \to v_H$ in $H^2(\Omega)^2$. Then $\partial_3 \varphi^n_3 = -\partial_1 \varphi^n_1 - \partial_2 \varphi^n_2$ is a Cauchy sequence in $H^1(\Omega)$, and by vertical Poincaré inequality, φ^n_3 is also a Cauchy sequence in $H^1(\Omega)$. Therefore, φ^n_3, being a Cauchy sequence in $H^1(\partial_3, \Omega)$, converges to a function v_3, which provides the desired lifting function. The continuous dependence (6.3) results from the above construction.

Now we take this $\mathbf{v} = (v_H, v_3)$ as a test function in (6.2) and integrate over $(t, t+h)$; i.e.,

$$(\tau_h u_H(t) - u_H(t), v_H) + \epsilon^2\, (\tau_h u_3(t) - u_3(t), v_3) = \int_t^{t+h} g^\epsilon(s)\, \mathrm{ds}, \tag{6.4}$$

where

$$\begin{aligned} g^\epsilon(s) \ \ = \ \ & (u_H, (\mathbf{u} \cdot \nabla) v_H) - \epsilon^2\, (\mathbf{u} \cdot \nabla u_3, v_3) - (b(u_H), v_H) - (\nabla_\nu u_H, \nabla_\nu v_H) \\ & - \epsilon\, (\nabla_\nu(\epsilon u_3), \nabla_\nu v_3) - \epsilon\, \{(\beta u_3, v_1) - (\beta u_1, v_3)\} + \langle \theta_H, v_H \rangle_{\Gamma_s}. \end{aligned}$$

Now we prove that

$$\|g^\epsilon\|_{L^{4/3}(0,T)} \le C\|v_H\|_{Y_H}. \tag{6.5}$$

To this end, we estimate every piece of g^ϵ. For the nonlinear terms, we have

$$(u_H, (\mathbf{u} \cdot \nabla) v_H) \le \|u_H\|_{L^3} \|\mathbf{u}\|_{L^2} \|\nabla v_H\|_{L^6} \le C\, \|u_H\|_{L^3} \|\mathbf{u}\|_{L^2} \|v_H\|_{Y_H}$$

and

$$\epsilon^2\, (\mathbf{u} \cdot \nabla u_3, v_3) \le \|\epsilon \mathbf{u}\|_{L^3} \|\nabla(\epsilon u_3)\|_{L^2} \|v_3\|_{L^6} \le C\, \|\epsilon \mathbf{u}\|_{L^3} \|\epsilon u_3\|_{H^1} \|v_H\|_{Y_H}.$$

By interpolation between $L^\infty(0, T; L^2)$ and $L^2(0, T; L^6)$, u_H is bounded in $L^4(0, T; L^3)$; i.e., $\|u_H\|_{L^3}$ is bounded in $L^4(0, T)$. As $\|\mathbf{u}\|_{L^2}$ is bounded in $L^2(0, T)$, we have $(u_H, (\mathbf{u} \cdot \nabla) v_H)$ bounded in $L^{4/3}(0, T)$. Similarly, as $\|\epsilon \mathbf{u}\|_{L^3}$ is bounded in $L^4(0, T)$ and $\|\epsilon u_3\|_{H^1}$ is bounded in $L^2(0, T)$, we have $\epsilon^2\, (\mathbf{u} \cdot \nabla u_3, v_3)$ bounded in $L^{4/3}(0, T)$. The linear terms of g^ϵ are handled easily by the Cauchy–Schwarz inequality:

$$\begin{array}{rcll} (b(u_H), v_H) & \le & \|u_H\|_{L^2} \|v_H\|_{L^2} & \text{bounded in } L^\infty(0, T), \\ (\nabla_\nu u_H, \nabla_\nu v_H) & \le & \|u_H\|_{H^1} \|v_H\|_{H^1} & \text{bounded in } L^2(0, T), \\ \epsilon\, (\nabla_\nu(\epsilon u_3), \nabla_\nu v_3) & \le & \epsilon\, \|\epsilon u_3\|_{H^1} \|v_3\|_{H^1} & \text{bounded in } L^2(0, T), \\ \epsilon\beta\, \{(u_3, v_1) - (u_1, v_3)\} & \le & 2f \|\epsilon \mathbf{u}\|_{L^2} \|\mathbf{v}\|_{L^2} & \text{bounded in } L^\infty(0, T), \\ \langle \theta_H, v_H \rangle_{\Gamma_s} & \le & C\, \|\theta_H\|_{H^{-1/2}(\Gamma_s)} \|v_H\|_{H^1} & \text{bounded in } L^2(0, T). \end{array}$$

Therefore, taking into account (6.3), according to all previous bounds, (6.5) holds. Next, applying the Hölder inequality to (6.5), we see that

$$\int_t^{t+h} |g^\epsilon(s)|\,\mathrm{ds} \le C\, h^{1/4} \|v_H\|_{Y_H}.$$

On the other hand,

$$|\epsilon^2\,(\tau_h u_3(t) - u_3(t), v_3)| \le \epsilon\,\{\|\tau_h(\epsilon u_3)(t)\|_{L^2} + \|\epsilon u_3(t)\|_{L^2}\}\|v_3\|_{L^2} \le \epsilon\, C\,\|v_H\|_{Y_H}$$

by virtue of Proposition 6.1.

These last two estimates together with (6.4) yield the required result. □

6.3. Convergence. Here we come back to the notation $\mathbf{u}^\epsilon$. The space-time weak form of the anisotropic Navier–Stokes equations (2.7)–(2.13) is as follows.

Find $\mathbf{u}^\epsilon = (u_H^\epsilon, u_3^\epsilon) \in L^2(0,T;\mathbf{V}) \cap L^\infty(0,T;L^2(\Omega)^3)$ such that

$$\begin{aligned}
&\int_0^T -(u_H^\epsilon, \partial_t v_H) - (u_H^\epsilon, (\mathbf{u}^\epsilon\cdot\nabla)v_H) + (b(u_H^\epsilon), v_H) + (\nabla_\nu u_H^\epsilon, \nabla_\nu v_H)\\
&+\epsilon^2\int_0^T -(u_3^\epsilon, \partial_t v_3) + (\mathbf{u}^\epsilon\cdot\nabla u_3^\epsilon, v_3) + (\nabla_\nu u_3^\epsilon, \nabla_\nu v_3)\\
&+\epsilon\,\beta\int_0^T (u_3^\epsilon, v_1) - (u_1^\epsilon, v_3) = -(u_{0H}, v_H(0)) - \epsilon^2(u_{03}, v_3(0)) + \int_0^T \langle\theta_H, v_H\rangle_{\Gamma_s}\\
&\forall\,\mathbf{v} = (v_H, v_3) \in H^1(0,T;\mathbf{V}),\ \text{with}\ \mathbf{v}(T) = 0.
\end{aligned}\tag{6.6}$$

The purpose of the following is to take the limit as $\epsilon \to 0$ in (6.6) to come to (4.1). By Propositions 6.1 and 6.2, it follows that $\mathbf{u}^\epsilon$ is bounded in $L^2(0,T;\mathbf{W})$ and u_H^ϵ is bounded in $L^\infty(0,T;B_H)$, allowing us to extract a subsequence, still denoted by $\mathbf{u}^\epsilon$, such that

$$\mathbf{u}^\epsilon = (u_H^\epsilon, u_3^\epsilon) \rightharpoonup \mathbf{u} = (u_H, u_3) \quad \text{in}\, L^2(0,T;\mathbf{W})\,\text{weak},$$

$$u_H^\epsilon \overset{\star}{\rightharpoonup} u_H \quad \text{in}\, L^\infty(0,T;B_H)\,\text{weak} - \star.$$

These weak convergences are enough to take the limit in the linear terms of (6.6) (cf. [1]). In particular, the terms of $\mathrm{O}(\epsilon)$ associated with the Coriolis acceleration vanish as ϵ tends to zero. Indeed,

$$\epsilon\,\beta\int_0^T (u_3^\epsilon, v_1) - (u_1^\epsilon, v_3) \le \epsilon\,2f\int_0^T \|\mathbf{u}^\epsilon\|_{L^2}\|\mathbf{v}\|_{L^2}$$

$$\le\ \epsilon\,2f\,\|\mathbf{u}^\epsilon\|_{L^2(0,T;L^2)}\|\mathbf{v}\|_{L^2(0,T;L^2)}\ \le\ \epsilon\,C\,\|\mathbf{v}\|_{L^2(0,T;L^2)}\ \le\ C\,\epsilon.$$

On the other hand, combining (6.1), Proposition 6.1, and Lemma 6.3, we can apply Theorem 5.1 for $p=\infty$ and the spaces $B_H \overset{compact}{\hookrightarrow} W_H' \hookrightarrow Y_H'$. Therefore, there exists a subsequence $u_H^\epsilon \to u_H$ in $\mathcal{C}(0,T;W_H')$ strong. Thus we get the weak time-continuity $u_H \in \mathcal{C}(0,T;W_H')$, so that the initial condition (2.20) makes sense for the horizontal velocities. On the other hand, the term of $0(\epsilon^2)$ related to the initial condition for the vertical velocity vanishes as ϵ tends to zero. Indeed,

$$-\epsilon^2(u_{03}, v_3(0)) \le \epsilon^2\|u_{03}\|_{L^2}\|v_3(0)\|_{L^2} \le \epsilon^2\|\mathbf{u}_0\|_{L^2}\|v_3\|_{\mathcal{C}(0,T;L^2)}\ \le\ C\,\epsilon^2.\tag{6.7}$$

Now the nonlinear terms fall into four types:

$$\text{(I)}\quad \epsilon^2\int_0^T (u_i^\epsilon \partial_i u_3^\epsilon, v_3)\,\mathrm{dt},\quad 1\le i\le 2,$$

$$\text{(II)}\quad \epsilon^2\int_0^T (u_3^\epsilon \partial_3 u_3^\epsilon, v_3)\,\mathrm{dt},$$

$$\text{(III)}\quad \int_0^T (u_i^\epsilon u_j^\epsilon, \partial_j v_i)\,\mathrm{dt},\quad 1\le i,\ j\le 2,$$

$$\text{(IV)}\quad \int_0^T (u_i^\epsilon u_3^\epsilon, \partial_3 v_i)\,\mathrm{dt},\quad 1\le i\le 2.$$

Type (I) term:

$$\epsilon^2\int_0^T (u_i^\epsilon \partial_i u_3^\epsilon, v_3)\,\mathrm{dt} \le \epsilon\int_0^T \|u_i^\epsilon\|_{L^6}\|\partial_i(\epsilon u_3^\epsilon)\|_{L^2}\|v_3\|_{L^3}$$

$$\le C\,\epsilon\,\|u_i^\epsilon\|_{L^2(0,T;H^1)}\|\partial_i(\epsilon u_3^\epsilon)\|_{L^2(0,T;L^2)}\|v_3\|_{\mathcal{C}(0,T;H^1)} \le C\,\epsilon.$$

Type (II) term:

$$\epsilon^2\int_0^T (u_3^\epsilon \partial_3 u_3^\epsilon, v_3)\,\mathrm{dt} \le \epsilon\int_0^T \|\epsilon u_3^\epsilon\|_{L^6}\|\partial_3 u_3^\epsilon\|_{L^2}\|v_3\|_{L^3}$$

$$\le C\,\epsilon\,\|\epsilon u_3^\epsilon\|_{L^2(0,T;H^1)}\|\partial_3 u_3^\epsilon\|_{L^2(0,T;L^2)}\|v_3\|_{\mathcal{C}(0,T;H^1)} \le C\,\epsilon.$$

Consequently, the type (I) and (II) terms are $O(\epsilon)$ and vanish as ϵ tends to zero.

To handle the type (III) and (IV) terms, we need some strong convergences. From compactness by perturbation (Theorem 5.1 for $p = 2$ and the spaces $W_H \overset{compact}{\hookrightarrow} B_H \hookrightarrow Y'_H$), there exists a subsequence, still denoted by u_H^ϵ, such that

$$u_H^\epsilon \to u_H \quad \text{in}\, L^2(0,T;L^2(\Omega)^2) \equiv L^2((0,T)\times\Omega)^2\,\text{strong.}$$

By Proposition 6.1, we have u_H^ϵ bounded in $L^\infty(0,T;L^2(\Omega)^2)\cap L^2(0,T;L^6(\Omega)^2)$, which by interpolation ensures that

$$u_H^\epsilon \,\text{is bounded in}\, L^{10/3}(0,T;L^{10/3}(\Omega)^2) \equiv L^{10/3}((0,T)\times\Omega)^2.$$

By the interpolation inequality again, for all $q : 2 \le q < 10/3$ there exists $\alpha : 0 < \alpha \le 1$ such that

$$\|u_H^\epsilon - u_H\|_{L^q} \le \|u_H^\epsilon - u_H\|_{L^2}^{\alpha}\|u_H^\epsilon - u_H\|_{L^{10/3}}^{1-\alpha}.$$

Thus

$$u_H^\epsilon \to u_H \quad \text{in}\, L^q((0,T)\times\Omega)\,\text{strong} \quad \forall\, q : 2\le q < 10/3. \tag{6.8}$$

Type (III) term: By the Hölder inequality and (6.8), we have

$$u_i^\epsilon u_j^\epsilon \to u_i u_j \quad \text{in}\, L^r((0,T)\times\Omega)\,\text{strong} \quad \forall\, r : 1\le r < 5/3$$

and for all $i, j = 1, 2$. On the other hand, by interpolation between $L^\infty(0,T;L^2)$ and $L^2(0,T;L^6)$, u_i^ϵ (and u_j^ϵ) is bounded in $L^{8/3}(0,T;L^4)$, and hence $u_i^\epsilon u_j^\epsilon$ is bounded in $L^{4/3}(0,T;L^2)$, and finally,

$$u_i^\epsilon u_j^\epsilon \rightharpoonup u_i u_j \quad \text{in}\, L^{4/3}(0,T;L^2)\,\text{weak}, \quad 1 \le i, \quad j \le 2.$$

In particular, we get

$$\int_0^T (u_i^\epsilon u_j^\epsilon, \partial_j v_i) \to \int_0^T (u_i u_j, \partial_j v_i), \quad 1 \le i, \quad j \le 2.$$

Indeed, $v_i \in \mathcal{C}(0,T;H^1)$ so that

$$\partial_j v_i \in L^\infty(0,T,L^2) \subset L^4(0,T,L^2) \equiv (L^{4/3}(0,T,L^2))'.$$

Type (IV) term: We have

$$u_3^\epsilon \rightharpoonup u_3 \quad \text{in}\, L^2(0,T;L^2)\,\text{weak}.$$

So by the Hölder inequality and (6.8),

$$u_i^\epsilon u_3^\epsilon \rightharpoonup u_i u_3 \quad \text{in}\, L^s(0,T;L^s)\,\text{weak} \quad \forall\, s : 1 \le s < 5/4$$

and for all $i = 1, 2$. On the other hand, it is easy to see that $u_i^\epsilon u_3^\epsilon$ is bounded in $L^{8/7}(0,T;L^{4/3})$, and hence

$$u_i^\epsilon u_3^\epsilon \rightharpoonup u_i u_3 \quad \text{in}\, L^{8/7}(0,T;L^{4/3})\,\text{weak}, \quad 1 \le i \le 2.$$

Now we shall have to slightly increase the regularity of the test functions of (4.1) to finish the limit process in the Type (IV) terms. For instance, assuming the additional regularity for the test functions $\partial_3 v_i \in L^8(0,T;L^4)$, we get

$$\int_0^T (u_i^\epsilon u_3^\epsilon, \partial_3 v_i) \to \int_0^T (u_i u_3, \partial_3 v_i), \quad 1 \le i \le 2.$$

In conclusion, the limit function $\mathbf{u}$ is a solution of the variational formulation (4.1) for all $\mathbf{v} = (v_H, v_3) \in H^1(0,T,\mathbf{V})$ with $v_H(T) = 0$ and $\partial_3 v_H \in L^8(0,T;L^4)$. Finally, we can argue by density, taking advantage of the regularity of each term of (4.1), and obtain that (4.1) holds for all $\mathbf{v} = (v_H, v_3) \in H^1(0,T,\mathbf{W})$ with $v_H(T) = 0$ and $\partial_3 v_H \in L^\infty(0,T;L^3)$; hence the proof of Theorem 3.1 is finished.

7. Concluding remarks.

7.1. Convergence of the pressure. By using the De Rham lemma [18] in the formulation (6.6) (resp., (4.1)), we can recover the potentials p^ϵ (resp., p) as distributions

$$\nabla p^\epsilon = \begin{pmatrix} -\partial_t u_1^\epsilon - \mathbf{u}^\epsilon \cdot \nabla u_1^\epsilon + \Delta_\nu u_1^\epsilon + \alpha u_2^\epsilon - \epsilon\beta u_3^\epsilon \\ -\partial_t u_2^\epsilon - \mathbf{u}^\epsilon \cdot \nabla u_2^\epsilon + \Delta_\nu u_2^\epsilon - \alpha u_1^\epsilon \\ -\epsilon^2 \left\{\partial_t u_3^\epsilon + \mathbf{u}^\epsilon \cdot \nabla u_3^\epsilon - \Delta_\nu u_3^\epsilon\right\} + \epsilon\beta u_1^\epsilon \end{pmatrix}, \tag{7.1}$$

respectively,

$$\nabla p = \begin{pmatrix} -\partial_t u_1 - \mathbf{u} \cdot \nabla u_1 + \Delta_\nu u_1 + \alpha u_2 \\ -\partial_t u_2 - \mathbf{u} \cdot \nabla u_2 + \Delta_\nu u_2 - \alpha u_1 \\ 0 \end{pmatrix}. \tag{7.2}$$

Moreover, (7.1) is also verified in $H^{-1}(0,T;H^{-1}(\Omega)^3)$ (i.e., in the dual space of $H^1_0(0,T;H^1_0(\Omega)^3)$), whereas (7.2) holds in $H^{-1}(0,T;W^{-1,3/2}(\Omega)^3)$ (i.e., in the dual space of $H^1_0(0,T;W^{1,3}_0(\Omega)^3)$). Proceeding as in subsection 6.3, we may derive that

$$\partial_i p^\epsilon \overset{\star}{\rightharpoonup} \partial_i p \quad \text{in } H^{-1}(0,T;W^{-1,3/2}(\Omega)), \quad i=1,2,$$

and

$$\|\partial_3 p^\epsilon\|_{H^{-1}(0,T;H^{-1}(\Omega))} \leq C\epsilon.$$

In particular, we have the strong convergence of $\partial_3 p^\epsilon$ to $\partial_3 p$.

Remark. The strong convergence of $\partial_3 p^\epsilon$ takes place in a better space than the weak convergence of $\partial_i p^\epsilon$, $i=1,2$. In some sense, this means that the validity of the hydrostatic approximation is less demanding than the validity of the horizontal momentum equations.

Remark. The above convergences can be slightly improved with respect to time. They remain true, replacing the space $H^{-1}(0,T;W^{-1,3/2}(\Omega))$ (resp., $H^{-1}(0,T;H^{-1}(\Omega))$) with the space $W^{-1,\infty}(0,T;W^{-1,3/2}(\Omega))$ (resp., $W^{-1,\infty}(0,T;H^{-1}(\Omega))$).

7.2. Orders of magnitude of the vertical velocity in the original domain.

The purpose of this last subsection is to interpret the previous results in the original domain Ω_ϵ. Consequently, we are going to consider $\mathbf{v}=(v_1,v_2,v_3)$, the weak solution in Ω_ϵ of problem (2.1)–(2.5), related to $\mathbf{u}^\epsilon=(u_1^\epsilon,u_2^\epsilon,u_3^\epsilon)$, a weak solution of problem (2.7)–(2.13) in Ω; see (2.6). First, it is important to notice that the true vertical velocity $v_3=\epsilon u_3^\epsilon$ is small with respect to the horizontal velocities v_i, $i=1,2$. Indeed, taking into account the estimates in Proposition 6.1 for u_i^ϵ, $i=1,2$, and Proposition 6.2 for u_3^ϵ, scaling off Ω to Ω_ϵ, we obtain

$$\frac{\|v_3\|_{L^2(0,T;L^2(\Omega_\epsilon))}}{\|v_i\|_{L^2(0,T;L^2(\Omega_\epsilon))}} = 0(\epsilon), \quad i=1,2.$$

By the same argument, we obtain

$$\frac{\|\partial_z v_3\|_{L^2(0,T;L^2(\Omega_\epsilon))}}{\|\partial_z v_i\|_{L^2(0,T;L^2(\Omega_\epsilon))}} = 0(\epsilon), \quad i=1,2.$$

This phenomenon is actually observed in most geophysical flows, which, therefore, are quasi-horizontal. It is striking that the vertical velocity goes to zero even if the initial vertical velocity is not assumed to be small. Looking at (6.7) in the proof of convergence, we need only that $\epsilon^2\|u_{03}\|_{L^2(\Omega)} \to 0$, that is, $\|v_{03}\|_{L^2(\Omega_\epsilon)}/\|v_{0i}\|_{L^2(\Omega_\epsilon)} = O(\epsilon^{-\alpha})$, $\alpha<1$.

Whereas, for the horizontal gradient, we cannot avail ourselves of Proposition 6.2, and we obtain only

$$\frac{\|\nabla_{x,y}\, v_3\|_{L^2(0,T;L^2(\Omega_\epsilon))}}{\|\nabla_{x,y}\, v_i\|_{L^2(0,T;L^2(\Omega_\epsilon))}} = 0(1), \quad i=1,2.$$

REFERENCES

[1] P. Azérad, *Analyse des équations de Navier-Stokes en bassin peu profond et de l'équation de transport*, Thèse de Doctorat ès sciences, Université de Neuchâtel, Neuchâtel, Switzerland, 1996; also available online from http://gala.univ-perp.fr/˜azerad.

[2] P. Azérad and F. Guillén, *Equations de Navier-Stokes en bassin peu profond: hydrostatique l'approximation*, C. R. Acad. Sci. Paris Sér. I Math., 329 (1999), pp. 961–966.

[3] P. Azérad, O. Besson, and F. Guillén, *Fluid flow in shallow domains: Mathematical analysis and numerical simulation*, in Proceedings of the IV Catalan Days of Applied Mathematics, C. Garcia, C. Olivé, and M. Sanroma, eds., Universitat Rovira i Virgili, Tarragona, Spain, 1998, pp. 9–16.

[4] O. Besson, M. R. Laydi, and R. Touzani, *Un modèle asymptotique en océanographie*, C. R. Acad. Sci. Paris Sér. I Math., 310 (1990), pp. 661–665.

[5] O. Besson and M. R. Laydi, *Some estimates for the anisotropic Navier-Stokes equations and for the hydrostatic approximation*, M2AN Math. Model. Numer. Anal., 26 (1992), pp. 855–865.

[6] D. Bresch, J. Lemoine, and J. Simon, *A vertical diffusion model for lakes*, SIAM J. Math. Anal., 30 (1999), pp. 603–622.

[7] D. Iftimie, *The* 3*D Navier-Stokes equations seen as a perturbation of the* 2*D Navier-Stokes equations*, Bull. Soc. Math. France, 127 (1999), pp. 473–517.

[8] O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York, 1969.

[9] R. Lewandowski, *Analyse Mathématique et Océanographie*, Masson, Paris, 1997.

[10] J.-L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.

[11] J.-L. Lions, R. Temam, and S. Wang, *New formulations of the primitive equations of the atmosphere and applications*, Nonlinearity, 5 (1992), pp. 237–288.

[12] J.-L. Lions, R. Temam, and S. Wang, *On the equations of large scale ocean*, Nonlinearity, 5 (1992), pp. 1007–1053.

[13] S. Montgomery-Smith, *Global regularity of the Navier Stokes equation on thin three-dimensional domains with periodic boundary conditions*, Electron. J. Differential Equations, 1999 (1999), pp. 1–19; also available online from http://ejde.math.unt.edu.

[14] B. Pascal, *De l'équilibre des liqueurs, Paris,* 1663, in Oeuvres complètes, coll. La Pléiade, Gallimard, Paris, 1998.

[15] J. Pedlosky, *Geophysical Fluid Dynamics*, Springer-Verlag, New York, 1987.

[16] G. Raugel and G. Sell, *Navier-Stokes equations in thin* 3*D domains* I: *Global attractors and global regularity of solutions*, J. Amer. Math. Soc., 6 (1993) pp. 503–568.

[17] J. Simon, *Compact sets in* $L^p(0,T;B)$, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–97.

[18] R. Temam, *Navier-Stokes Equations*, North-Holland, Elsevier, Amsterdam, 1985.

[19] R. Temam, *Sur la stabilité et la convergence de la méthode des pas fractionnaires*, Ann. Mat. Pura Appl. (4), 79 (1968), pp. 191–379.

[20] R. Temam and M. Ziane, *Navier-Stokes equations in three-dimensional thin domains with various boundary conditions*, Adv. Differential Equations, 1 (1996), pp. 499–546.

[21] R. Temam and M. Ziane, *Navier-Stokes equations in thin spherical domains*, in Optimization Methods in Partial Differential Equations, Contemp. Math. 209, AMS, Providence, RI, 1997, pp. 281–314.

[22] R. K. Zeytounian, *Modélisation asymptotique en mécanique des fluides newtoniens*, Springer-Verlag, New York, 1994.

SIAM J. Math. Anal.
Vol. 33, No. 4, pp. 860–876

A MATHEMATICAL MODEL OF THE WEARING PROCESS OF A NONCONVEX STONE*

HITOSHI ISHII† AND TOSHIO MIKAMI‡

Abstract. We formulate the wearing process of a nonconvex stone in terms of partial differential equations (PDEs). We establish a comparison theorem, an existence theorem, and some stability properties of solutions of this PDE.

Key words. geometric evolution equations, wearing process, nonconvex stone

AMS subject classifications. 53C44, 35K65

PII. S0036141001384958

1. Introduction.

In this paper we study a mathematical model of the wearing process of a stone rolling on beach.

In [6], Firey proposed a mathematical model of the wearing process of such a stone in the case when it has a convex shape. In his model, a stone evolves according to the Gauss curvature flow. See, for instance, [1, 3, 7] for the mathematical developments regarding the Gauss curvature flow.

We extend his arguments to the case when the stone does not necessarily have a convex shape. The idea of this extension is very simple, and it is explained as follows. Let $V_t \subset \mathbf{R}^{n+1}$ be a stone at time t being worn due to hits of the bottom of the sea (or beach). In Firey's and our model the bottom of the sea is supposed to be a hyperplane. We fix a coordinate system for which the stone does not rotate and translate, and, for each unit vector $p \in \mathbf{R}^{n+1}$, we associate the hyperplane

$$H_p = \{x \in \mathbf{R}^{n+1} \mid x \cdot p = 0\}.$$

The set V_t evolves by losing its volume near the point where it is hit by the hyperplane H_p. Here the meaning of the sentence "A point $Q \in V_t$ is hit by H_p" is that the half space

$$Q + \{x \in \mathbf{R}^{n+1} \mid x \cdot p > 0\}$$

does not intersect with V_t. Of course, if $Q \in V_t$ is hit by a hyperplane, then $Q \in \partial V_t$. Three assumptions in this model of this wearing process are as follows: (i) the probability of H_p hitting the stone V_t is uniform in the direction p; (ii) the volume loss near a point $Q \in \partial V_t$ is proportional to how often the point Q is hit by hyperplanes; and (iii) the total volume loss of the stone in a time period is proportional to the length of the time period. Of course, once V_t becomes empty at a time t_0, then, by definition, $V_t = \emptyset$ for all $t > t_0$.

*Received by the editors February 13, 2001; accepted for publication August 13, 2001; published electronically December 28, 2001.

http://www.siam.org/journals/sima/33-4/38495.html

†Department of Mathematics, School of Education, Nishi-Waseda 1–6–1, Shinjuku-ku, Tokyo 169–8050, Japan (ishii@edu.waseda.ac.jp). This author was supported in part by the Grant-in-Aid for Scientific Research, 12440044, 123D4406, JSPS.

‡Department of Mathematics, Hokkaido University, Sapporo 060–0810, Japan (mikami@math.sci.hokudai.ac.jp). This author was supported in part by the Grant-in-Aid for Scientific Research, 13640096, JSPS.

In what follows, we restrict our study to the case when the boundary of the stone V_t at time t is given by the graph of an evolving function. That is, we study the case when V_t is given by

$$V_t = \{(x, y) \in \mathbf{R}^n \times \mathbf{R} \mid y \geq u(x, t)\}$$

for some function $u : \mathbf{R}^n \times [0, \infty) \to \mathbf{R}$. Restricting our study to this case is not realistic in applications to the wearing process of a stone, but we are so far forced to do so by technical difficulties. Despite this restriction, we believe that the results obtained here are of some interest at least from the mathematical viewpoint. A natural approach to a PDE model of the wearing process of a compact stone seems to be the level set approach. We refer to [2] for this approach to the Gauss curvature flow.

In the case of a convex stone, the PDE which the function u should satisfy is

$$u_t(x, t) = g(Du(x, t), D^2u(x, t)) \qquad \text{for } (x, t) \in \mathbf{R}^n \times (0, \infty),$$

where $g : \mathbf{R}^n \times \mathcal{S}^n \to \mathbf{R}$ is given by

$$g(p, X) = \frac{\det_+ X}{(1 + |p|^2)^{(n+1)/2}}.$$

Here and later we denote by $\mathcal{S}^n$ the space of all real $n \times n$ symmetric matrices, and, for $X \in \mathcal{S}^n$,

$$\det{}_+ X = \prod_{i=1}^{n} \max\{\lambda_i, 0\},$$

with λ_i denoting the eigenvalues of X.

The extension to the general case is straightforward, and, in the general case, the PDE for u to satisfy is

$$\text{(1.1)} \quad u_t(x, t) = \chi(u, Du(x, t), x, t)g(Du(x, t), D^2u(x, t)) \qquad \text{for } (x, t) \in \mathbf{R}^n \times (0, \infty),$$

where $\chi : \mathcal{F} \times \mathbf{R}^n \times \mathbf{R}^n \times (0, \infty) \to \{0, 1\}$ is given by

$$\chi(u, p, x, t) = \begin{cases} 1 & \text{if } u(y, t) \geq u(x, t) + p \cdot (y - x) \quad \text{for } y \in \mathbf{R}^n, \\ 0 & \text{otherwise,} \end{cases}$$

and $\mathcal{F}$ denotes the space of all real functions on $\mathbf{R}^n \times (0, \infty)$. The heuristic meaning of χ in (1.1) is that if $\chi(u, p, x, t) = 0$, then the point $(x, u(x, t))$ on the surface (of the stone) $y \geq u(x, t)$ is not hit by the hyperplane (the sea bottom) $H_{(p,-1)}$, and otherwise it is hit by $H_{(p,-1)}$. The precise meaning of (1.1) will be clarified in the next section.

One of the new features in the PDE (1.1) is its nonlocality due to the factor χ.

Our primary purposes are to establish a comparison result for solutions of (1.1) and an existence theorem of solutions of (1.1). We use the notion of viscosity solution adapted to (1.1), and some stability results are established as well. The comparison result is stated and proved in section 2, the existence result is treated in section 3, and stability properties of viscosity solutions of (1.1) are discussed in section 4.

Finally, we wish to pursue some properties of solutions of (1.1) in a future publication.

2. Comparison theorem. We first introduce a function which describes the asymptotic behavior of solutions we shall be concerned with. Let h_0 be a real-valued function on $\mathbf{R}^n$. Assume that

$$h_0 \in C(\mathbf{R}^n) \quad \text{and} \quad h_0(x) \geq \varepsilon_0 |x| \quad \forall x \in \mathbf{R}^n, \tag{2.1}$$

for some constant $\varepsilon_0 > 0$.

Let $u, v : \mathbf{R}^n \times [0, \infty) \to \mathbf{R}$. We make the following assumptions:

$$u \in \mathrm{USC}(\mathbf{R}^n \times [0, \infty)),\ v \in \mathrm{LSC}(\mathbf{R}^n \times [0, \infty)), \tag{2.2}$$

$$u(x, 0) \leq v(x, 0)\ \forall x \in \mathbf{R}^n. \tag{2.3}$$

For each $T > 0$,

$$\sup_{(x,t) \in \mathbf{R}^n \times [0,T]} (|u(x, t) - h_0(x)| + |v(x, t) - h_0(x)|) < \infty. \tag{2.4}$$

u satisfies

$$u_t(x, t) \leq \chi^+(u, Du(x, t), x, t) g(Du(x, t), D^2 u(x, t)) \quad \text{in } \mathbf{R}^n \times (0, \infty) \tag{2.5}$$

in the viscosity sense, where

$$\chi^+(u, p, x, t) = \chi(u, p, x, t).$$

To be more precise, we call $u \in \mathrm{USC}(\mathbf{R}^n \times (0, \infty))$ a viscosity subsolution of (1.1) if whenever $\varphi \in C^2(\mathbf{R}^n \times (0, \infty))$ and $u - \varphi$ attains its maximum at $(\hat{x}, \hat{t}) \in \mathbf{R}^n \times (0, \infty)$, then

$$\varphi_t(\hat{x}, \hat{t}) \leq \chi^+(u, D\varphi(\hat{x}, \hat{t}), \hat{x}, \hat{t}) g(D\varphi(\hat{x}, \hat{t}), D^2\varphi(\hat{x}, \hat{t})).$$

The condition (2.5) is now stated that u is a viscosity subsolution of (1.1). v satisfies

$$v_t(x, t) \geq \chi^-(v, Dv(x, t), x, t) g(Dv(x, t), D^2 v(x, t)) \quad \text{in } \mathbf{R}^n \times (0, \infty) \tag{2.6}$$

in the viscosity sense, where

$$\chi^- \equiv \chi^-(v, p, x, t) = 1$$

if

$$v(y, t) > v(x, t) + p \cdot (y - x) \quad \text{for } y \in \mathbf{R}^n \setminus \{x\},$$

and there is $\varepsilon > 0$ such that for all $(y, s) \in \mathbf{R}^n \times (0, \infty)$ satisfying $|y| > \varepsilon^{-1}$ and $|s - t| < \varepsilon$,

$$v(y, s) > p \cdot y + \varepsilon |y|,$$

and, otherwise,

$$\chi^- \equiv \chi^-(v, p, x, t) = 0.$$

Again, to be precise, we call a function $v \in \mathrm{LSC}(\mathbf{R}^n \times (0,\infty))$ a viscosity supersolution of (1.1) if whenever $\varphi \in C^2(\mathbf{R}^n \times (0,\infty))$ and $u-\varphi$ attains its minimum at $(\hat{x},\hat{t}) \in \mathbf{R}^n \times (0,\infty)$, then

$$\varphi_t(\hat{x},\hat{t}) \geq \chi^-(u, D\varphi(\hat{x},\hat{t}), \hat{x}, \hat{t}) g(D\varphi(\hat{x},\hat{t}), D^2\varphi(\hat{x},\hat{t})).$$

The exact meaning of (2.6) is that v is a viscosity supersolution of (1.1).

We call a function $u : \mathbf{R}^n \times (0,\infty) \to \mathbf{R}$ a viscosity solution of (1.1) if the function

$$u^*(x,t) := \lim_{r\downarrow 0} \sup\{u(y,s) \mid |y-x| + |s-t| \leq r\}$$

is a viscosity subsolution of (1.1) and the function

$$u_*(x,t) := \lim_{r\downarrow 0} \inf\{u(y,s) \mid |y-x| + |s-t| \leq r\}$$

is a viscosity supersolution of (1.1). We often suppress "viscosity" for the sake of simplicity of presentation.

THEOREM 1. *Assume that* (2.1)–(2.6) *hold.* (a) *For any $\theta \in (0,1)$, the inequality $u(x,\theta t) \leq v(x,t)$ holds for all $(x,t) \in \mathbf{R}^n \times [0,\infty)$.* (b) *Assume, in addition, that*

$$h_0 \in C^2(\mathbf{R}^n),$$

$$\det{}_+ D^2 h_0(x) \leq C\left(1 + |Dh_0(x)|^2\right)^{(n+1)/2} \quad \forall x \in \mathbf{R}^n,$$

for some constant $C > 0$ and that for each $\varepsilon > 0$ there is a constant $R \equiv R(\varepsilon) > 0$ such that for all $x \in \mathbf{R}^n$, if $|x| \geq R$, then

$$u(x,0) - \varepsilon \leq h_0(x) \leq v(x,0) + \varepsilon.$$

Then $u \leq v$ on $\mathbf{R}^n \times [0,\infty)$.

Remark. Note that under the assumptions of (b) above, the function

$$w(x,t) = Ct + h_0(x)$$

satisfies

$$w_t(x,t) \geq g(Dw(x,t), D^2 w(x,t)) \quad \forall (x,t) \in \mathbf{R}^n \times (0,\infty).$$

Proof. First, we prove part (a). We fix $\theta \in (0,1)$ and $T > 0$ and intend to prove that

$$u(x,\theta t) \leq v(x,t) \quad \forall (x,t) \in \mathbf{R}^n \times [0,T). \tag{2.7}$$

Define

$$u_\theta(x,t) := u(x,\theta t) \quad \forall (x,t) \in \mathbf{R}^n \times [0,\infty),$$

and observe that

$$u_{\theta,t} \leq \theta \chi^+ g(Du_\theta, D^2 u_\theta) \quad \text{in } \mathbf{R}^n \times [0,\infty).$$

In view of (2.4), we may assume, if necessary, by adding u and v a constant that $u \geq 0$ on $\mathbf{R}^n \times [0,T)$.

Let $\mu \in (0,1)$ be such that $\theta\mu^{1-n} \leq 1$. Noting that for $(p,x,t) \in \mathbf{R}^n \times \mathbf{R}^n \times (0,\infty)$, if

$$u_\theta(y,t) \geq u_\theta(x,t) + p \cdot (y-x) \quad \forall y \in \mathbf{R}^n,$$

then

$$\mu u_\theta(y,t) \geq \mu u_\theta(x,t) + \mu p \cdot (y-x) \quad \forall y \in \mathbf{R}^n,$$

and computing formally that, since $\theta\mu \leq \mu^n$,

$$\begin{aligned}\mu u_{\theta,t} &\leq \mu\theta u_t(x,\theta t) \leq \mu^n \chi^+(u_\theta, Du_\theta, x, t)\frac{\det_+ D^2 u_\theta}{(1+|Du_\theta|^2)^{(n+1)/2}} \\ &\leq \chi^+(\mu u_\theta, \mu Du_\theta, x, t)\frac{\det_+(\mu D^2 u_\theta)}{(1+|\mu Du_\theta|^2)^{(n+1)/2}},\end{aligned}$$

we see that μu_θ is a subsolution of

$$u_t \leq \chi^+ g(Du, D^2 u) \quad \text{ in } \mathbf{R}^n \times (0,\infty).$$

In order to prove (2.7), it is enough to show that for all $\mu \in (0,1)$ satisfying $\theta\mu^{1-n} \leq 1$, we have

$$\mu u_\theta \leq v \quad \text{ on } \mathbf{R}^n \times [0,T). \tag{2.8}$$

Fix $\mu \in (0,1)$ such that $\theta\mu^{1-n} \leq 1$. To see that (2.8) holds for this μ, we suppose to the contrary that

$$\sup_{\mathbf{R}^n \times [0,T)} (\mu u_\theta - v) > 0.$$

By assumption (2.4), there is a constant $C_0 > 0$ such that

$$|u(x,t) - h_0(x)| + |v(x,t) - h_0(x)| \leq C_0 \quad \forall (x,t) \in \mathbf{R}^n \times [0,T].$$

From this, we see that for all $x \in \mathbf{R}^n$ and $t, s \in [0,T]$,

$$\begin{aligned}\mu u_\theta(x,t) \leq\ & \mu(h_0(x) + C_0) \leq v(x,s) + C_0 + (\mu - 1)(h_0(x) + C_0) \\ \leq\ & v(x,s) - (1-\mu)\varepsilon_0|x| + 2C_0.\end{aligned} \tag{2.9}$$

In particular, we see that there is $R > 0$ such that, for all $(x,t) \in \mathbf{R}^n \times [0,T]$, if $|x| \geq R$, then

$$\mu u_\theta(x,t) \leq v(x,t).$$

For notational simplicity, we write $w := \mu u_\theta$.

Subtracting from w a function like

$$\frac{\alpha}{T + \alpha^2 - t},$$

where α is a small positive constant, we may assume that

$$w_t \leq \chi^+ g(Dw, D^2 w) - \delta \quad \forall (x,t) \in \mathbf{R}^n \times (0,T),$$

for some constant $\delta > 0$. Moreover, there is a constant $\gamma \in (0, T/2)$ such that

$$w \le v \quad \forall (x,t) \in B(0,R) \times ([0,\gamma] \cup [T-\gamma, T].$$

Set

$$Q := B(0, R+\gamma) \times [0,T]$$

and

$$M := \sup_Q (|w| + |v|).$$

We assume for the moment that for some $(\xi, \tau) \in Q$ and $p \in \mathbf{R}^n$,

$$w(x,\tau) \ge w(\xi,\tau) + p \cdot (x - \xi) \quad \forall x \in \mathbf{R}^n. \tag{2.10}$$

Then we observe by using (2.9) that for any $(\eta, \sigma) \in Q$,

$$\begin{aligned} v(x,\sigma) &\ge w(x,\tau) + (1-\mu)\varepsilon_0|x| - 2C_0 \ge w(\xi,\tau) + p\cdot(x-\xi) + (1-\mu)\varepsilon_0|x| - 2C_0 \\ &= v(\eta,\sigma) + p\cdot(x-\eta) + (1-\mu)\varepsilon_0|x| + p\cdot(\eta-\xi) + w(\xi,\tau) - v(\eta,\sigma) - 2C_0. \end{aligned}$$

Observe as well that

$$h_0(x) + C_0 \ge h_0(\xi) - C_0 + p \cdot (x-\xi) \quad \forall x \in \mathbf{R}^n$$

and hence that

$$|p| \le 2C_0 + h_0(\xi) - h_0(\xi + |p|^{-1}p) \le 2C_0 + 2 \sup_{B(0,R+\gamma+1)} |h_0|.$$

Now we choose a constant $L > 0$ so that

$$\frac{1}{2}(1-\mu)\varepsilon_0(R+L) > 2C_0 + 4(R+\gamma)\left(C_0 + \sup_{B(0,R+\gamma)} |h_0|\right) + M,$$

and, consequently, for all $x \in \mathbf{R}^n$ and $(\eta,\sigma) \in Q$, if $|x| \ge R + L$ and if (2.10) holds for some $(\xi,\tau) \in Q$ and $p \in \mathbf{R}^n$, then

$$v(x,\sigma) > v(\eta,\sigma) + p \cdot (x - \eta). \tag{2.11}$$

Set $Q_L := B(0, R+L+\gamma) \times [0,T]$. Let $\varepsilon \in (0,1)$ and

$$w^\varepsilon(x,t) = \max\left\{ w(y,s) - \frac{1}{2\varepsilon}(|x-y|^2 + |t-s|^2) \;\middle|\; (y,s) \in Q_L \right\},$$

and

$$v_\varepsilon(x,t) = \min\left\{ v(y,s) + \frac{1}{2\varepsilon}(|x-y|^2 + |t-s|^2) \;\middle|\; (y,s) \in Q_L \right\}.$$

It is well known that

$$w \le w^\varepsilon \quad \text{and} \quad v \ge v_\varepsilon \quad \text{on } Q_L$$

and that if $\varepsilon > 0$ is sufficiently small, then

$$w^\varepsilon(x,t) > w(y,s) - \frac{1}{2\varepsilon}(|x-y|^2 + |t-s|^2)$$

and

$$v_\varepsilon(x,t) < v(y,s) + \frac{1}{2\varepsilon}(|x-y|^2 + |t-s|^2)$$

hold for all $(x,t),(y,s) \in Q_L$ such that $|x-y|^2 + |t-s|^2 \geq (\gamma/2)^2$. Fix such a small $\varepsilon > 0$. Set $Q' := \mathrm{int}B(0, R+\gamma/2) \times (\gamma/2, T-\gamma/2)$. Again, it is well known that if $(x,t) \in Q'$ and $(p,q,X) \in J^{2,+}w^\varepsilon(x,t)$, then, for $y = x + \varepsilon p$ and $s = t + \varepsilon q$,

$$w^\varepsilon(x,t) = w(y,s) - \frac{1}{2\varepsilon}\left(|x-y|^2 + |t-s|^2\right)$$

and

$$(p,q,X) \in J^{2,+}w(y,s).$$

We refer to [4] for the definition of $J^{2,\pm}$. Similarly, if $(x,t) \in Q'$ and $(p,q,X) \in J^{2,-}v_\varepsilon(x,t)$, then, for $y = x - \varepsilon p$ and $s = t - \varepsilon q$,

$$v_\varepsilon(x,t) = v(y,s) + \frac{1}{2\varepsilon}\left(|x-y|^2 + |t-s|^2\right)$$

and

$$(p,q,X) \in J^{2,-}v(y,s).$$

Finally, as is well known, w^ε and $-v_\varepsilon$ are Lipschitz continuous and semiconvex on $\mathbf{R}^{n+1}$.

We maximize functions which are small perturbations of the function $w^\varepsilon - v_\varepsilon$ on Q_L. First, we note that

$$\max_{Q_L \setminus Q'}(w^\varepsilon - v_\varepsilon) \leq 0 < \max_{Q_L}(w^\varepsilon - v_\varepsilon). \tag{2.12}$$

By perturbed optimization techniques (see, e.g., [5, Corollary 3.8]), for each $n \in \mathbf{N}$ there is $(a_n, b_n) \in \mathbf{R}^n \times \mathbf{R}$ such that the function

$$w^\varepsilon(x,t) - v_\varepsilon(x,t) - a_n \cdot x - b_n t$$

on Q_L attains a strict maximum at a point $(x_n, t_n) \in Q_L$ and such that $|a_n| + |b_n| < 1/n$.

Focusing our attention on large $n \in \mathbf{N}$, in view of (2.12), we may assume that $(x_n, t_n) \in Q'$.

For each such $n \in \mathbf{N}$ there is a function $\psi_n \in C^2(\mathbf{R}^n)$ such that

$$\psi_n(x) > 0 \quad \forall x \neq x_n, \qquad \psi_n(x_n) = 0;$$

$$\psi_n(x) \text{ is strictly convex on } \mathbf{R}^n;$$

$$w^\varepsilon(x,t) - v_\varepsilon(x,t) - a_n \cdot x - b_n t + \psi_n(x) \text{ attains a strict maximum at } (x_n, t_n);$$

$$\|\psi_n\|_\infty + \|D\psi_n\|_\infty + \|D^2\psi_n\|_\infty < 1/n.$$

By Jensen's maximum principle (see, e.g., [4, Lemmas A.2 and A.3]), there are $(p_n, q_n) \in \mathbf{R}^n \times \mathbf{R}$ and $(y_n, s_n) \in Q_L$ such that $w^\varepsilon(x,t) - v_\varepsilon(x,t) + \psi_n(x) - (a_n + p_n)\cdot x - (b_n + q_n)t$ attains a maximum at (y_n, s_n) over Q_L, w^ε and v_ε are twice differentiable at (y_n, s_n), and $|p_n| + |q_n| < 1/n$. We may assume that $(y_n, s_n) \in Q'$.

By the elementary maximum principle, we have

$$P_n := Dw^\varepsilon(y_n, s_n) = Dv_\varepsilon(y_n, s_n) - D\psi_n(y_n) + a_n + p_n,$$

$$Q_n := w^\varepsilon_t(y_n, s_n) = v_{\varepsilon,t}(y_n, s_n) + b_n + q_n,$$

$$X_n := D^2 w^\varepsilon(y_n, s_n) \le D^2 v_\varepsilon(y_n, s_n) - D^2\psi_n(y_n).$$

By the semiconvexity of w^ε and $-v_\varepsilon$, we have

$$-\frac{1}{\varepsilon} I \le D^2 w^\varepsilon(y_n, s_n), \quad D^2 v_\varepsilon(y_n, s_n) \le \frac{1}{\varepsilon} I.$$

Moreover, w^ε and v_ε are Lipschitz continuous on Q. For example, there is a constant $C \equiv C(\varepsilon) > 0$ such that for $(x,t), (y,s) \in Q$,

$$|w^\varepsilon(x,t) - w^\varepsilon(y,s)| + |v_\varepsilon(x,t) - v_\varepsilon(y,s)| \le C(|x-y| + |t-s|).$$

As a consequence, we have

$$|Dw^\varepsilon(y_n, s_n)| + |Dv_\varepsilon(y_n, s_n)| + |w^\varepsilon_t(y_n, s_n)| + |v_{\varepsilon,t}(y_n, s_n)| \le 4C.$$

Let $(\xi_n, t_n) \in Q_L$ and $(\eta_n, \sigma_n) \in Q_L$ be the maximum and minimum points of functions

$$w(x,t) - \frac{1}{2\varepsilon}\left(|x - y_n|^2 + |t - s_n|^2\right)$$

and

$$v(x,t) + \frac{1}{2\varepsilon}\left(|x - y_n|^2 + |t - s_n|^2\right)$$

on Q_L, respectively. Recall that

$$|\xi_n - y_n|^2 + |\tau_n - \sigma_n|^2 < (\gamma/2)^2 \quad \text{and} \quad |\eta_n - y_n|^2 + |\sigma_n - s_n|^2 < (\gamma/2)^2$$

and hence that $(\xi_n, \tau_n),\ (\eta_n, \sigma_n) \in \text{int}Q$.

Now we have

$$Q_n \le \chi^+(w, P_n, \xi_n, \tau_n) g(P_n, X_n) - \delta;$$

$$\begin{aligned} Q_n - q_n - b_n \ge & \chi^-(v, P_n + D\psi_n(y_n) - a_n - p_n, \eta_n, \sigma_n) \\ & \cdot g(P_n + D\psi_n(y_n) - a_n - p_n, X_n + \psi_n(y_n)). \end{aligned}$$

If $\chi^+(w, P_n, \xi_n, \tau_n) = 0$, then

$$Q_n + \delta \le 0 \le Q_n - q_n - b_n. \tag{2.13}$$

Suppose instead that $\chi^+(w, P_n, \xi_n, \tau_n) = 1$. This yields

$$w(x, \tau_n) \geq w(\xi_n, \tau_n) + P_n \cdot (x - \xi_n) \quad \forall x \in \mathbf{R}^n. \tag{2.14}$$

We have from this

$$w(x, \tau_n) - \frac{1}{2\varepsilon}(|y_n - \xi_n|^2 + |s_n - \tau_n|^2) \geq w^\varepsilon(y_n, s_n) + P_n \cdot (x - \xi_n) \quad (x \in \mathbf{R}^n).$$

If $y \in B(0, R + L + \gamma/2)$, then $x := y + \xi_n - y_n \in B(0, R + L + \gamma)$, and we have

$$w^\varepsilon(y_n, s_n) + P_n \cdot (y - y_n) \leq w(y + \xi_n - y_n, \tau_n) - \frac{1}{2\varepsilon}(|y_n - \xi_n|^2 + |s_n - \tau_n|^2) \leq w^\varepsilon(y, s_n).$$

Since (y_n, s_n) is a maximum point of

$$w^\varepsilon(x, t) - v_\varepsilon(x, t) + \psi_n(x) - (a_n + p_n) \cdot x - (q_n + b_n)t$$

over Q_L, we have for any $x \in \mathbf{R}^n$, with $|x| \leq R + L + \gamma$,

$$\begin{aligned}
&w^\varepsilon(x, s_n) - v_\varepsilon(x, s_n) + \psi_n(x) - (a_n + p_n) \cdot x - (q_n + b_n)s_n \\
&\leq w^\varepsilon(y_n, s_n) - v_\varepsilon(y_n, s_n) + \psi_n(y_n) - (a_n + p_n) \cdot y_n - (q_n + b_n)s_n.
\end{aligned}$$

Thus, using the strict convexity of ψ_n, for any $x \neq y_n$ with $|x| \leq R + L + \gamma/2$, we have

$$\begin{aligned}
v_\varepsilon(x, s_n) &\geq v_\varepsilon(y_n, s_n) + w^\varepsilon(x, s_n) - w^\varepsilon(y_n, s_n) + \psi_n(x) - \psi_n(y_n) \\
&\quad - (a_n + p_n) \cdot (x - y_n) \\
&\geq v^\varepsilon(y_n, s_n) + (P_n - a_n - p_n) \cdot (x - y_n) + \psi_n(x) - \psi_n(y_n) \\
&> v_\varepsilon(y_n, s_n) + (P_n - a_n - p_n + D\psi_n(y_n)) \cdot (x - y_n).
\end{aligned}$$

Therefore, we have, for all $x \in B(0, R + L + \gamma/2)$ with $x \neq y_n$ and $(y, s) \in Q_L$,

$$\begin{aligned}
v(y, s) &+ \frac{1}{2\varepsilon}\left(|y - x|^2 + |s - s_n|^2\right) \\
&> v(\eta_n, \sigma_n) + \frac{1}{2\varepsilon}\left(|\eta_n - y_n|^2 + |\sigma_n - s_n|^2\right) \\
&\quad + (P_n - a_n - p_n + D\psi_n(y_n)) \cdot (x - y_n),
\end{aligned}$$

and hence, for $y \in B(0, R + L)$ with $y \neq \eta_n$, plugging $s = \sigma_n$, $x = y + y_n - \eta_n \in B(0, R + L + \gamma/2) \setminus \{y_n\}$, we get

$$v(y, \sigma_n) > v(\eta_n, \sigma_n) + (P_n - a_n - p_n + D\psi_n(y_n)) \cdot (y - \eta_n).$$

On the other hand, by our choice of L (see (2.11)), we have from (2.14)

$$v(y, \sigma_n) > v(\eta_n, \sigma_n) + P_n \cdot (y - \eta) + \frac{1}{2}(1 - \mu)\varepsilon_0|y| \tag{2.15}$$

for all $y \in \mathbf{R}^n$ satisfying $|y| \geq R + L$. Choosing n large enough, we may assume that for all $y \in \mathbf{R}^n$, if $|y| \geq R + L$, then

$$(|a_n| + |p_n| + |D\psi_n(y_n)|)(|y| + R + \gamma) \leq \frac{1}{2}(1 - \mu)\varepsilon_0|y|.$$

Then we have

$$v(x,\sigma_n) > v(\eta_n,\sigma_n) + (P_n - a_n - p_n + D\psi_n(y_n)) \cdot (x - \eta_n) \quad \forall x \in \mathbf{R}^n \setminus \{\eta_n\}.$$

This, together with (2.15), guarantees that

$$\chi^-(v, P_n + D\psi_n(y_n) - a_n - p_n, \eta_n, \sigma_n) = 1,$$

and we have

$$g(P_n + D\psi_n(y_n) - a_n - p_n, X_n + \psi_n(y_n)) \le Q_n - q_n - b_n \le g(P_n, X_n) - \delta.$$

This and (2.13) yield a contradiction as we send $n \to \infty$.

Next we prove part (b). Fix $\varepsilon > 0$. Select $R > 0$ so that

$$u(x,0) - \varepsilon \le h_0(x) \le v(x,0) + \varepsilon \quad \forall x \in \mathbf{R}^n \setminus B(0,R).$$

We note that there is a function $k_\varepsilon(x) \in C(B(0,R+1))$ such that

$$u(x,0) - \varepsilon/2 \le k_\varepsilon(x) \le v(x,0) + \varepsilon/2 \quad \forall x \in B(0,R+1). \tag{2.16}$$

Indeed, since $u(x,0)$ is upper semicontinuous on $B(0,R+1)$, we find a sequence of continuous functions j_n on $B(0,R+1)$ such that

$$j_n(x) \downarrow u(x,0) \quad \text{as } n \to \infty \quad \forall x \in B(0,R+1).$$

Then, noting that

$$(j_n(x) - v(x,0))_+ \downarrow 0 \quad \text{as } n \to \infty$$

and the functions $(j_n(x) - v(x,0))_+$ are upper semicontinuous on $B(0,R+1)$, by virtue of Dini's lemma, we see that

$$(j_n(x) - v(x,0))_+ \downarrow 0$$

uniformly on $B(0,R+1)$ as $n \to \infty$. Selecting n large enough, the function j_n satisfies (2.16) with $k_\varepsilon = j_n$. Now a simple argument based on the mollification of k_ε and on partition of unity, we see that there is a function $h_\varepsilon \in C^2(\mathbf{R}^n)$ such that

$$u(x,0) - \varepsilon \le h_\varepsilon(x) \le v(x,0) + \varepsilon \quad \forall x \in \mathbf{R}^n,$$

and, for some constant $C_\varepsilon > 0$,

$$\det{}_+ D^2 h_\varepsilon(x) \le C_\varepsilon (1 + |Dh_\varepsilon(x)|^2)^{(n+1)/2} \quad \forall x \in \mathbf{R}^n. \tag{2.17}$$

Set $z(x,t) := h_\varepsilon(x) + \varepsilon + C_\varepsilon t$ for $(x,t) \in \mathbf{R}^n \times [0,\infty)$. By (2.17) we see that z is a supersolution of

$$z_t \ge g(Dz, D^2 z) \quad \text{in } \mathbf{R}^n \times (0,\infty).$$

Moreover, z satisfies not only this property but also the other properties required for v in part (a). Hence we see that, for any $\theta \in (0,1)$,

$$u(x,\theta t) \le z(x,t) \quad \forall (x,t) \in \mathbf{R}^n \times [0,\infty),$$

i.e.,

$$u(x,t) \le z(x,\theta^{-1}t) \quad \forall (x,t) \in \mathbf{R}^n \times [0,\infty).$$

Since z is continuous, we get

$$u(x,t) \le z(x,t) \quad \forall (x,t) \in \mathbf{R}^n \times [0,\infty).$$

Now let $\delta > 0$, and observe that

$$u(x,\delta) \le z(x,\delta) = h_\varepsilon(x) + \varepsilon + C_\varepsilon\delta \le v(x,0) + 2\varepsilon + C_\varepsilon\delta \quad \forall x \in \mathbf{R}^n.$$

Define

$$w(x,t) := u(x,\delta+t) - 2\varepsilon - C_\varepsilon\delta \quad \forall (x,t) \in \mathbf{R}^n \times [0,\infty).$$

Then w satisfies all the properties required for u in part (a), and hence we have

$$w(x,\theta t) \le v(x,t) \quad \forall (x,t) \in \mathbf{R}^n \times [0,\infty),\ \forall \theta \in (0,1).$$

For example, we have

$$u(x,\delta+\theta t) \le v(x,t) + 2\varepsilon + C_\varepsilon\delta \quad ((x,t) \in \mathbf{R}^n \times [0,\infty)\ \forall \theta \in (0,1).$$

Therefore, for all $(x,t) \in \mathbf{R}^n \times (\delta,\infty)$, choosing $\theta = (t-\delta)/t \in (0,1)$ in the above, we have

$$u(x,t) \le v(x,t) + 2\varepsilon + C_\varepsilon\delta.$$

Letting $\delta \downarrow 0$, we see that $u(x,t) \le v(x,t) + 2\varepsilon$ for all $(x,t) \in \mathbf{R}^n \times (0,\infty)$. It is immediate to conclude that $u \le v$ in $\mathbf{R}^n \times [0,\infty)$. □

3. Existence theorem. Let $h \in C(\mathbf{R}^n)$ satisfy

$$\lim_{|x|\to\infty} \big|h(x) - h_0(x)\big| = 0, \tag{3.1}$$

where $h_0 \in C^2(\mathbf{R}^n)$ is a function satisfying (2.1) and

$$\det{}_+ D^2 h_0(x) \le C(1 + |Dh_0(x)|^2)^{(n+1)/2} \quad \forall x \in \mathbf{R}^n, \tag{3.2}$$

for some constant $C > 0$.

THEOREM 2. *Assume that* (3.1) *and* (3.2) *hold. Then there is a viscosity solution* $u \in C(\mathbf{R}^n \times [0,\infty))$ *of*

$$\begin{cases} u_t(x,t) = \chi(u, Du(x,t), x, t) g(Du(x,t), D^2u(x,t)) & \textit{in } \mathbf{R}^n \times (0,\infty), \\ u(x,0) = h(x) \qquad \textit{for } x \in \mathbf{R}^n \end{cases} \tag{3.3}$$

satisfying

$$\sup_{(x,t)\in\mathbf{R}^n\times[0,T]} |u(x,t) - h(x)| < \infty \quad \forall T > 0. \tag{3.4}$$

Proof of Theorem 2. *Step* 1. *Construction of a supersolution:* Let $\varepsilon \in (0,1)$, and let $h_\varepsilon \in C^2(\mathbf{R}^n)$ be a function such that

$$|h(x) - h_\varepsilon(x)| < \varepsilon \quad \forall x \in \mathbf{R}^n;$$

$$\det{}_+ D^2 h_\varepsilon(x) \le C_\varepsilon (1 + |Dh_\varepsilon(x)|^2)^{(n+1)/2} \quad \forall x \in \mathbf{R}^n,$$

for some constant $C_\varepsilon > 0$.

Set

$$w(x,t;\varepsilon) := \varepsilon + h_\varepsilon(x) + C_\varepsilon t \quad \forall (x,t) \in \mathbf{R}^n \times [0,\infty).$$

Then the function $w(x,t;\varepsilon)$ of (x,t) is a supersolution of

$$w_t \ge \chi^- g(Dw, D^2 w) \quad \text{in } \mathbf{R}^n \times (0,\infty)$$

and satisfies

$$h(x) \le w(x,t;\varepsilon) \le h(x) + 2\varepsilon + C_\varepsilon t \quad \forall (x,t) \in \mathbf{R}^n \times [0,\infty). \tag{3.5}$$

Define

$$v(x,t) := \inf_{\varepsilon \in (0,1)} w(x,t;\varepsilon) \quad \forall (x,t) \in \mathbf{R}^n \times [0,\infty).$$

Since the pointwise infimum of supersolutions is a supersolution, the function v_* is a supersolution of

$$v_t \ge \chi^- g(Dv, D^2 v) \quad \text{in } \mathbf{R}^n \times (0,\infty).$$

Moreover, v is USC on $\mathbf{R}^n \times [0,\infty)$, $v(x,0) = h(x)$ for all $x \in \mathbf{R}^n$ by (3.5), and, for each $T > 0$, $h(x) \le v(x,t) \le h(x) + C_T$ for all $(x,t) \in \mathbf{R}^n \times [0,T]$ and for some constant $C_T > 0$ by (3.5).

Step 2. *Construction of a subsolution.* We set

$$z(x,t) = h(x) \quad \forall (x,t) \in \mathbf{R}^n \times [0,\infty).$$

Then z is a subsolution of

$$z_t \le \chi^+ g(Dz, D^2 z) \quad \forall (x,t) \in \mathbf{R}^n \times (0,\infty)$$

and satisfies

$$z(x,0) = h(x) \quad \forall x \in \mathbf{R}^n;$$

$$z(x,t) \le h(x) \quad \forall x \in \mathbf{R}^n, \ \forall t \ge 0.$$

Step 3. *Perron's method.* We apply Perron's method (see Theorem 5) to obtain a solution u of (3.3) such that

$$h(x) \le u(x,t) \le v(x,t) \quad \forall (x,t) \in \mathbf{R}^n \times [0,\infty).$$

This inequality shows that

$$\sup_{(x,t) \in \mathbf{R}^n \times [0,T]} |u(x,t) - h(x)| < \infty \quad \forall T > 0$$

and that

$$\lim_{t \downarrow 0} \sup_{x \in \mathbf{R}^n} |u(x,t) - h(x)| = 0.$$

Now Theorem 1 yields that $u^* \le u_*$ on $\mathbf{R}^n \times [0,\infty)$ and therefore that $u \in C(\mathbf{R}^n \times [0,\infty))$. □

4. Stability properties. In this section, we establish some stability results for viscosity solutions of (1.1), the nature of which is rather standard in the theory of viscosity solutions [4]. Note that Theorem 5 has already been applied in the proof of Theorem 2.

Theorem 3. *Let* $\{u^\alpha\}$ *be a family of viscosity subsolutions of* (1.1). *Assume that the function*

$$u(x,t) := \Big(\sup_\alpha u^\alpha\Big)^*(x,t)$$

on $\mathbf{R}^n \times (0,\infty)$ *is locally bounded. Then* u *is a viscosity subsolution of* (1.1).

Theorem 4. *Let* $\{u^\alpha\}$ *be a family of viscosity supersolutions of* (1.1). *Assume that the function*

$$u(x,t) := \Big(\inf_\alpha u^\alpha\Big)_*(x,t)$$

on $\mathbf{R}^n \times (0,\infty)$ *is locally bounded. Then* u *is a viscosity supersolution of* (1.1).

Theorem 5. *Let* $f^- \in \mathrm{LSC}(\mathbf{R}^n \times [0,\infty))$ *and* $f^+ \in \mathrm{USC}(\mathbf{R}^n \times [0,\infty))$ *be a viscosity subsolution and a viscosity supersolution of* (1.1)*, respectively. Suppose that* $f^- \le f^+$ *on* $\mathbf{R}^n \times [0,\infty)$. *Let* $\mathcal{S}$ *be the set of viscosity subsolutions* v*, defined on* $\mathbf{R}^n \times [0,\infty)$*, of* (1.1) *such that* $f^- \le v \le f^+$ *on* $\mathbf{R}^n \times [0,\infty)$*, and set*

$$u(x,t) := \sup\{v(x,t) \mid v \in \mathcal{S}\} \quad \textit{for } (x,t) \in \mathbf{R}^n \times [0,\infty).$$

Then u *is a viscosity solution of* (1.1).

Proof of Theorem 3. Let $\varphi \in C^2(\mathbf{R}^n \times (0,\infty))$. Let $(\hat{x},\hat{t}) \in \mathbf{R}^n \times (0,\infty)$ be a maximum point of $u - \varphi$. We may assume that $(\hat{x},\hat{t})$ is a strict maximum point and that u^α is USC in $\mathbf{R}^n \times (0,\infty)$.

As usual, we can select sequences of α_n and of points $(x_n,t_n) \in \mathbf{R}^n \times (0,\infty)$ so that $u^{\alpha_n} - \varphi$ attains a local maximum at (x_n,t_n) and so that, as $n \to \infty$,

$$u^{\alpha_n}(x_n,t_n) \to u(\hat{x},\hat{t}) \quad \text{and} \quad (x_n,t_n) \to (\hat{x},\hat{t}).$$

In view of standard proofs, we need only to show that if

$$\chi^+\left(u^{\alpha_n}, D\varphi(x_n,t_n), x_n, t_n\right) = 1 \quad \forall n \in \mathbf{N}, \tag{4.1}$$

then

$$\chi^+(u, D\varphi(\hat{x},\hat{t}), \hat{x}, \hat{t}) = 1.$$

Let us assume that (4.1) holds. Then, by definition, we have

$$u^{\alpha_n}(y,t_n) \ge u^{\alpha_n}(x_n,t_n) + D\varphi(x_n,t_n)\cdot(y - x_n) \quad \forall y \in \mathbf{R}^n.$$

Hence we have

$$u(y,t_n) \ge u^{\alpha_n}(x_n,t_n) + D\varphi(x_n,t_n)\cdot(y - x_n) \quad \forall y \in \mathbf{R}^n.$$

Sending $n \to \infty$, we get

$$u(y,\hat{t}) \ge u(\hat{x},\hat{t}) + D\varphi(\hat{x},\hat{t})\cdot(y - \hat{x}) \quad \forall y \in \mathbf{R}^n,$$

i.e.,

$$\chi^+(u, D\varphi(\hat{x},\hat{t}), \hat{x},\hat{t}) = 1. \qquad \square$$

Proof of Theorem 4. Let $\varphi \in C^2(\mathbf{R}^n \times (0,\infty))$. Let $(\hat{x},\hat{t}) \in \mathbf{R}^n \times (0,\infty)$ be a minimum point of $u - \varphi$. We may assume that $(\hat{x},\hat{t})$ is a strict minimum point and that u^α is LSC in $\mathbf{R}^n \times (0,\infty)$.

As before, there are sequences of α_n and of points $(x_n, t_n) \in \mathbf{R}^n \times (0,\infty)$ so that $u^{\alpha_n} - \varphi$ attains a local minimum at (x_n, t_n) and, as $n \to \infty$,

$$u^{\alpha_n}(x_n,t_n) \to u(\hat{x},\hat{t}) \quad \text{and} \quad (x_n,t_n) \to (\hat{x},\hat{t}).$$

In view of standard proofs, we need only to show that if

$$\liminf_{n\to\infty} \chi^-\left(u^{\alpha_n}, D\varphi(x_n,t_n), x_n, t_n\right) = 0, \tag{4.2}$$

then

$$\chi^-(u, D\varphi(\hat{x},\hat{t}), \hat{x},\hat{t}) g(D\varphi(\hat{x},\hat{t}), D^2\varphi(\hat{x},\hat{t})) = 0. \tag{4.3}$$

We argue by contradiction and so assume that (4.2) holds and that (4.3) does not hold. By passing to a subsequence, we may assume that

$$\chi^-\left(u^{\alpha_n}, D\varphi(x_n,t_n), x_n, t_n\right) = 0 \quad \forall n \in \mathbf{N}. \tag{4.4}$$

We have

$$\chi^-(u, D\varphi(\hat{x},\hat{t}), \hat{x},\hat{t}) = 1; \tag{4.5}$$

$$g(D\varphi(\hat{x},\hat{t}), D^2\varphi(\hat{x},\hat{t})) > 0. \tag{4.6}$$

Then from (4.5) there is a constant $\varepsilon > 0$ such that, for $(x,t) \in \mathbf{R}^n \times (0,\infty)$, if $|x| > \varepsilon^{-1}$, $|t - \hat{t}| < \varepsilon$, then

$$u(x,t) > p \cdot x + 2\varepsilon|x|, \tag{4.7}$$

where $p := D\varphi(\hat{x},\hat{t})$. We write p_n for $D\varphi(x_n,t_n)$ as well. We may assume that $|t_n - \hat{t}| < \varepsilon/2$ and $|p_n - p| < \varepsilon$ for all $n \in \mathbf{N}$.

From (4.7), for all $n \in \mathbf{N}$ and $(x,t) \in \mathbf{R}^n \times (0,\infty)$, if $|x| > \varepsilon^{-1}$, $|t - t_n| < \varepsilon/2$, then we have

$$u^{\alpha_n}(x,t) > p_n \cdot x + \varepsilon|x|.$$

Therefore, we see from (4.4) that there is a sequence of points $y_n \in \mathbf{R}^n$ such that $y_n \neq x_n$ and such that

$$u^{\alpha_n}(y_n,t_n) \leq u^{\alpha_n}(x_n,t_n) + p_n \cdot (y_n - x_n), \tag{4.8}$$

which immediately yields

$$u(y_n,t_n) \leq u^{\alpha_n}(x_n,t_n) + p_n \cdot (y_n - x_n). \tag{4.9}$$

We see easily from (4.7) and (4.9) that there is a constant $R > 0$ such that

$$|y_n| \leq R \quad \forall n \in \mathbf{N}.$$

Thus there is a subsequence of $n \in \mathbf{N}$ such that along the subsequence, $\{y_n\}$ converges to a point $\hat{y} \in \mathbf{R}^n$. We have from (4.9)

$$u(\hat{y}, \hat{t}) \leq u(\hat{x}, \hat{t}) + p \cdot (\hat{y} - \hat{x}).$$

Now (4.6) implies that the matrix $D^2\varphi(\hat{x}, \hat{t})$ is positive definite. By continuity, there is a constant $\delta > 0$ such that

$$D^2\varphi(x,t) \geq \delta I \quad ((x,t) \in B(\hat{x}, \delta) \times [\hat{t} - \delta, \hat{t} + \delta]).$$

We may assume that $(x_n, t_n) \in B(\hat{x}, \delta/2) \times [\hat{t} - \delta/2, \hat{t} + \delta/2]$ for all $n \in \mathbf{N}$. By using Taylor's expansion, we see that, for all $n \in \mathbf{N}$ and $x \in B(x_n, \delta/2)$,

$$u^{\alpha_n}(x, t_n) \geq u^{\alpha_n}(x_n, t_n) + p_n \cdot (x - x_n) + \frac{\delta}{2}|x - x_n|^2.$$

This, together with (4.8), guarantees that $|y_n - x_n| > \delta/2$ for all $n \in \mathbf{N}$ and hence that $|\hat{y} - \hat{x}| \geq \delta/2$. Thus we see that

$$\chi^-(u, p, \hat{x}, \hat{t}) = 0,$$

which is a contradiction. □

Proof of Theorem 5. From Theorem 3, we know that u_* is a subsolution of

$$u_t \leq \chi^+ g(Du, D^2u) \quad \text{in } \mathbf{R}^n \times (0, \infty).$$

We must show that u_* is a supersolution of

$$u_t \geq \chi^- g(Du, D^2u) \quad \text{in } \mathbf{R}^n \times (0, \infty).$$

To do this, we argue by contradiction and so suppose that there are $\varphi \in C^2(\mathbf{R}^n \times (0, \infty))$ and $(\hat{x}, \hat{t}) \in \mathbf{R}^n \times (0, \infty)$ such that

$$\varphi_t(\hat{x}, \hat{t}) < \chi^-(u_*, D\varphi(\hat{x}, \hat{t}), \hat{x}, \hat{t}) g(D\varphi(\hat{x}, \hat{t}), D^2\varphi(\hat{x}, \hat{t}))$$

and such that $u_* - \varphi$ attains a strict minimum at $(\hat{x}, \hat{t})$. We may assume that $u_*(\hat{x}, \hat{t}) = \varphi(\hat{x}, \hat{t})$.

If

$$\chi^-(u_*, D\varphi(\hat{x}, \hat{t}), \hat{x}, \hat{t}) g(D\varphi(\hat{x}, \hat{t}), D^2\varphi(\hat{x}, \hat{t})) = 0,$$

then φ is a classical subsolution of

$$u_t < 0$$

in a neighborhood of $(\hat{x}, \hat{t})$, and the standard argument for Perron's method yields a contradiction.

Next consider the case where

$$\chi^-(u_*, D\varphi(\hat{x}, \hat{t}), \hat{x}, \hat{t}) g(D\varphi(\hat{x}, \hat{t}), D^2\varphi(\hat{x}, \hat{t})) > 0.$$

Let us write $\hat{u} = u_*(\hat{x}, \hat{t})$, $a = D\varphi(\hat{x}, \hat{t})$, and $b = \varphi_t(\hat{x}, \hat{t})$.

Fix $\eta > 0$ so that

$$D^2\varphi(x,t) > 0 \quad \forall (x,t) \in Q_\eta.$$

Here Q_η denotes the cylinder $\{(x,t) \in \mathbf{R}^n \times \mathbf{R} \mid |x-\hat{x}| < \eta,\ |t-\hat{t}| < \eta\}$, and η is chosen so small that $Q_\eta \subset \mathbf{R}^n \times (0,\infty)$.

Since

$$\chi^-(u_*, a, \hat{x}, \hat{t}) = 1,$$

we have

$$u_*(x,\hat{t}) > \hat{u} + a \cdot (x-\hat{x}) \quad \forall x \in \mathbf{R}^n \setminus \{\hat{x}\},$$

and

$$u_*(y,s) > a \cdot y + \delta|y| \quad \text{if } |y| > \delta^{-1},\ |s-\hat{t}| < \delta,$$

for some $\delta > 0$. From these, we deduce that there is $\gamma > 0$ such that if $|p-a| < \gamma$, $|t-\hat{t}| < \gamma$, and $|r-\hat{u}| < \gamma$, then

$$u_*(x,t) > r + p \cdot (x-\hat{x}) \quad \text{if } |x-\hat{x}| \geq \eta.$$

As usual in the proof of existence by Perron's method, we find $r > 0$ and $\varepsilon > 0$ such that

$$Q_{2r} \subset \mathbf{R}^n \times (0,\infty);$$

$$\varphi(x,t) + \varepsilon < u(x,t) \quad \text{if } (x,t) \in Q_{2r} \setminus \overline{Q}_r;$$

$$\varphi(x,t) + \varepsilon \leq g(x,t) \quad \text{if } (x,t) \in Q_{2r};$$

$$\varphi_t(x,t) < g(D\varphi(x,t), D^2\varphi(x,t)) \quad \text{if } (x,t) \in Q_{2r}.$$

We may assume that $Q_{2r} \subset Q_\eta$ and that

$$|\varphi(x,t) + \varepsilon - \hat{u}| < \gamma \quad \forall (x,t) \in Q_{2r},$$

$$|D\varphi(x,t) - a| < \gamma \quad \forall (x,t) \in Q_{2r},$$

and that $2r < \gamma$.

Note that for $(x,t) \in Q_{2r}$ and $y \in \mathbf{R}^n$, if $|y-\hat{x}| \geq \eta$, then we have

$$u_*(y,t) > \varphi(x,t) + \varepsilon + D\varphi(x,t) \cdot (y-x). \tag{4.10}$$

Define

$$v(x,t) := \begin{cases} u(x,t) & \text{for } (x,t) \in (\mathbf{R}^n \times (0,\infty)) \setminus \overline{Q}_r, \\ \max\{u(x,t), \varphi(x,t)+\varepsilon\} & \text{for } (x,t) \in \overline{Q}_r. \end{cases}$$

We intend to show that v^* is a subsolution in $\mathbf{R}^n \times (0,\infty)$. Let $\psi \in C^2(\mathbf{R}^n \times (0,\infty))$. Let $(y,s) \in \mathbf{R}^n \times (0,\infty)$ be a maximum point of the function $v^* - \psi$.

Noting that $v \geq u$ in $\mathbf{R}^n \times (0,\infty)$ and that $v = u$ on $\mathbf{R}^n \times (0,\infty) \setminus \overline{Q}_r$, we see that for $(p,x,t) \in \mathbf{R}^n \times \mathbf{R}^n \times (0,\infty)$, if $(x,t) \notin \overline{Q}_r$, then

$$\chi^+(v^*, p, x, t) \geq \chi^+(u^*, p, x, t).$$

Hence, if $(y,s) \notin \overline{Q}_r$, then

$$\psi_t(y,s) \le \chi^+(v^*, D\psi(y,s), y, s) g(D\psi(y,s), D^2\psi(y,s)).$$

Now assume that $(y,s) \in Q_{2r}$. If $v^*(y,s) = u^*(y,s)$, then, as above, we get

$$\psi_t(y,s) \le \chi^+(v^*, D\psi(y,s), y, s) g(D\psi(y,s), D^2\psi(y,s)).$$

If $v^*(y,s) = \varphi(y,s) + \varepsilon$, then $\varphi - \psi$ has a local maximum at (y,s). Hence we have

$$D\varphi(y,s) = D\psi(y,s), \quad \varphi_t(y,s) = \psi_t(y,s), \quad D^2\varphi(y,s) \le D^2\psi(y,s),$$

and, therefore,

$$\psi_t(y,s) \le g(D\psi(y,s), D^2\psi(y,s)).$$

Since $D^2\varphi(x,t) > 0$ in Q_{2r}, we have

$$\varphi(x,s) + \varepsilon \ge \varphi(y,s) + \varepsilon + D\varphi(y,s) \cdot (x-y) \quad \text{if } |x - \hat{x}| < 2r$$

and hence

$$v^*(x,s) \ge v^*(y,s) + D\varphi(y,s) \cdot (x-y) \quad \text{if } |x - \hat{x}| < 2r.$$

Furthermore, from (4.10) we have

$$v^*(x,s) \ge u^*(x,s) \ge \varphi(y,s) + \varepsilon + D\varphi(y,s) \cdot (x-y) \quad \text{if } x - \hat{x}| \ge 2r.$$

These together yield

$$v^*(x,s) \ge v^*(y,s) + D\psi(y,s) \cdot (x-y) \quad \text{if } x \in \mathbf{R}^n.$$

Thus we see that

$$\chi^+(v^*, D\psi(y,s), y, s) = 1$$

and conclude that

$$\psi_t(y,s) \le \chi^+(v^*, D\psi(y,s), y, s) g(D\psi(y,s), D^2\psi(y,s)),$$

which shows that v^* is a subsolution. Since $f^- \le v \le f^+$ and $v \not\le u$, this yields a contradiction. □

Acknowledgment. Hitoshi Ishii would like to thank the Department of Mathematics, Tokyo Metropolitan University, where most of this work was done.

REFERENCES

[1] B. Andrews, *Gauss curvature flow: The fate of the rolling stones*, Invent. Math., 138 (1999), pp. 151–161.

[2] D. Chopp, L. C. Evans, and H. Ishii, *Waiting time effects for Gauss curvature flows*, Indiana Univ. Math. J., 48 (1999), pp. 311–334.

[3] B. Chow, *Deforming convex hypersurfaces by the nth root of the Gaussian curvature*, J. Differential Geom., 22 (1985), pp. 117–138.

[4] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[5] I. Ekeland and G. Lebourg, *Generic Frechet-differentiability and perturbed optimization problems in Banach spaces*, Trans. Amer. Math. Soc., 224 (1976), pp. 193–216.

[6] W. J. Firey, *Shapes of worn stones*, Mathematika, 21 (1974), pp. 1–11.

[7] K. Tso, *Deforming a hypersurface by its Gauss-Kronecker curvature*, Comm. Pure Appl. Math., 38 (1985), pp. 67–882.

SIAM J. MATH. ANAL.
Vol. 33, No. 4, pp. 877–895

AN UP-TO-THE BOUNDARY VERSION OF FRIEDRICHS'S LEMMA AND APPLICATIONS TO THE LINEAR KOITER SHELL MODEL*

ADEL BLOUZA[†] AND HERVÉ LE DRET[‡]

Abstract. In this work, we introduce a variant of the standard mollifier technique that is valid up to the boundary of a Lipschitz domain in $\mathbb{R}^n$. A version of Friedrichs's lemma is derived that gives an estimate up to the boundary for the commutator of the multiplication by a Lipschitz function and the modified mollification. We use this version of Friedrichs's lemma to prove the density of smooth functions in the new function space introduced in our earlier work concerning the linear Koiter shell model for shells with little regularity. The density of smooth functions is in turn used to prove continuous dependence of the solution of Koiter's model on the midsurface. This provides a complete justification of our new formulation of the Koiter model.

Key words. Friedrichs's lemma, shell theory, Koiter's model

AMS subject classifications. 35A99, 74K25

PII. S0036141000380012

1. Introduction.

Mollification is a basic technique in analysis. It is classically performed by convolution with a compactly supported mollifier. In order for the convolution to be defined, it is necessary either to work on the whole of $\mathbb{R}^n$ or in a compactly contained subdomain ω of the domain of interest Ω. For many classical function spaces on a domain Ω, approximation by smooth functions is quite straightforward if the boundary of the domain Ω is regular enough. In effect, in the latter case, there usually is a continuous extension operator that reduces the case of the domain to that of $\mathbb{R}^n$. It suffices to perform the mollification on the extended function and then restrict the mollifed function to the domain. See [1], [13], [14], [22], and [24] among others.

Our main field of application here is the linear Koiter shell model in elasticity (see [21]). In [5], [6], we introduced a new formulation of this model that makes sense and is well-posed for midsurfaces of class $W^{2,\infty}$ instead of C^3, as was customarily assumed earlier; see also [9] and [11] for shell models in the same context of regularity. The simplest and most natural examples of $W^{2,\infty}$-shells are given by globally C^1- and piecewise C^3-midsurfaces. Consider, for instance, a shell consisting of a planar part that is connected to a circular cylinder part or an egg-shaped shell made of a quarter of a sphere and a quarter of an ellipsoid glued together along a circle. Our new formulation entails the introduction of a new functional setting. The new function space involves multiplication of distributional partial derivatives of the functions by given Lipschitz functions related to the geometry of the midsurface. It is required that such quantities be square-integrable. To the best of our knowledge, this specific kind of function space was not studied before regarding such fundamental issues as the density or nondensity of smooth functions. There are relatively close ideas in

*Received by the editors October 24, 2000; accepted for publication (in revised form) May 21, 2001; published electronically December 28, 2001.

http://www.siam.org/journals/sima/33-4/38001.html

[†] Laboratoire d'Analyse et Modélisation Stochastique, Université de Rouen, 76821 Mont-St-Aignan Cedex, France (Adel.Blouza@univ-rouen.fr).

[‡] Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie, Boîte Courrier 187, 75252 Paris Cedex 05, France (ledret@ccr.jussieu.fr). The work of this author was completed during a stay at the Texas Institute for Computational and Applied Mathematics, University of Texas at Austin, with the support of the TICAM Visiting Fellowship program.

transport equation theory (see [2], [12], [10]) although the techniques used therein do not apply in our case. For the function space introduced in [5], there is no obvious extension operator, and it is not natural to work on the whole of $\mathbb{R}^2$. Thus different ideas were needed to address the density question.

One such idea was first put forth in [23], without mention of regularity for the domain, and was then rediscovered simultaneously later on by the authors and by [15] in a slightly different form. The idea consists in defining a new mollification technique in which the mollifier is simultaneously scaled and translated inside the domain. For the technique to work, it is necessary that the domain satisfy a uniform cone condition, which is practically equivalent to being Lipschitz; see [1], [16]. This simple but powerful idea yields mollified functions defined on the whole domain without any need for an extension operator to provide values for the function outside of its domain of definition.

It is a straightforward matter to reproduce the proof of Friedrichs's lemma using the above convolution-translation in place of the convolution itself. This gives an L^p estimate on the whole domain for the commutator of the convolution-translation and the multiplication by a Lipschitz function applied to partial derivatives of an L^p function.

This version of Friedrichs's lemma is the main tool in proving the density of smooth functions in the new function space for Koiter's model. This density has important consequences. For example, it shows that standard finite element methods actually approximate the solution of the newly formulated variational problem for Koiter's model; see [19] and [20]. Another consequence that we develop here is that if we consider a sequence of $W^{2,\infty}$-midsurfaces that converge in a natural sense toward a given midsurface and a sequence of loads that also converge, then the corresponding sequence of solutions to Koiter's model converges in a natural sense too. Since the new model coincides with the classical model for C^3-midsurfaces, taking a sequence of such C^3-midsurfaces converging to a midsurface that is only $W^{2,\infty}$ shows that our new formulation is an appropriate extension of the classical formulation to less regular midsurfaces from the mathematical point of view.

2. A modified mollification technique.

In this section, we develop the convolution-translation technique introduced in [23] (see also [15]) that allows for up-to-the-boundary mollification without requiring an extension operator. It is well known that the density of smooth functions in, for example, Sobolev spaces, may fail if the domain under consideration is not regular. It is thus natural that the regularity of the boundary should come into play.

First of all, let us recall the *uniform cone property* for a domain Ω in $\mathbb{R}^n$. We refer the reader to [1], [13], [16], and [24] for details.

DEFINITION 2.1. *An open set $\Omega \subset \mathbb{R}^n$ is said to satisfy the* cone property *if there exists an open cone*

$$C = \{x = (x', x_n) \in \mathbb{R}^n; 0 < x_n < h, |x'| < x_n \tan(\theta/2)\},$$

with $h > 0$ and $0 < \theta < \pi$, such that for every point x in $\bar{\Omega}$ there is a rotation R_x such that $\bar{C}_x = x + R_x\bar{C} \subset \bar{\Omega}$. In other words, any point x is the vertex of a cone congruent to C and included in $\bar{\Omega}$ (or Ω). Here $|\cdot|$ denotes the standard Euclidean norm either on $\mathbb{R}^{n-1}$ or on $\mathbb{R}^n$.

The set Ω is said to satisfy the uniform cone property *if there exists a locally finite open covering $\{U_i\}_{i\geq 1}$ of $\partial\Omega$, and a corresponding sequence $\{C_i\}$ of cones, each congruent to some fixed cone C, such that the following hold.*

(i) *There exists M such that every U_i has diameter less than M.*

(ii) *For some $\delta > 0$, $\Omega_\delta = \{x \in \Omega; \operatorname{dist}(x; \partial\Omega) < \delta\} \subset \bigcup_{i=1}^{\infty} U_i$.*

(iii) *For every $i \in \mathbb{N}$, $Q_i = \bigcup_{x \in \Omega \cap U_i} (x + C_i) \subset \Omega$.*

(iv) *There exists $N \in \mathbb{N}$ such that every collection of $N + 1$ of the sets Q_i has an empty intersection.*

Remark 2.2. Note that a bounded domain of $\mathbb{R}^n$ satisfies the uniform cone property if and only if its boundary is Lipschitz; see [7] or [16, Theorem 1.2.2.2]. □

Let us now introduce some notation for the convolution-translation operator, which is the basic tool for subsequent developments.

DEFINITION 2.3. *Let e be a unit vector in $\mathbb{R}^n$ and $\tau > 0$. For all $u, v \in L^1(\mathbb{R}^n)$, we define their convolution-translation (of amount τ in the direction e) $u \star_{\tau,e} v$ by*

$$u \star_{\tau,e} v(x) = u \star v(x - \tau e), \tag{2.1}$$

that is,

$$u \star_{\tau,e} v(x) = \int_{\mathbb{R}^n} u(x - \tau e - y) v(y)\, dy. \tag{2.2}$$

Obviously, $u \star_{\tau,e} v \in L^1(\mathbb{R}^n)$, $u \star_{\tau,e} v = v \star_{\tau,e} u$, and if v is C^∞, so is $u \star_{\tau,e} v$ with $\partial^\alpha(u \star_{\tau,e} v) = u \star_{\tau,e} \partial^\alpha v$ for any multi-index $\alpha \in \mathbb{N}^n$.

The interesting feature of this slightly modified convolution is that it can be used to define a mollification technique for Sobolev spaces that is valid up to the boundary of any domain in $\mathbb{R}^n$ that satisfies the uniform cone condition, without using any extension operator. Let us now describe how this can be achieved.

Let ρ be a standard mollifier, i.e., a positive C^∞ function on $\mathbb{R}^n$ supported in the unit ball and such that $\int_{\mathbb{R}^n} \rho(x)\, dx = 1$. Let Ω be a domain in $\mathbb{R}^n$ that satisfies the uniform cone condition, and denote by U_i the locally finite covering of the boundary from Definition 2.1. To entirely cover Ω, we let $U_0 = \{x \in \Omega; \operatorname{dist}(x; \partial\Omega) > \delta/2\}$. We denote by $(\varphi_i)_{i \in \mathbb{N}}$ an associated C^∞ partition of unity.

THEOREM 2.4. *For all $u \in W^{m,p}(\Omega)$, there exists a sequence u_ε in $C^\infty(\bar{\Omega})$ such that*

$$u_\varepsilon \to u \quad \text{in} \quad W^{m,p}(\Omega) \quad \text{when} \quad \varepsilon \to 0. \tag{2.3}$$

Proof. This is of course a very classical result. We only include it here to show how it can be proved using the convolution-translation instead of standard mollification together with an extension operator.

We begin by localizing $u = \sum_{i \in \mathbb{N}} u_i$ with $u_i = \varphi_i u$. Each u_i has compact support in $U_i \cap \bar{\Omega}$. The "interior" part u_0 does not pose any problem and can be approximated by standard mollification. Let us concentrate on what happens near the boundary.

From now on, we may thus assume that u has support in, say, $U_1 \cap \bar{\Omega} = U$ without loss of generality. As far as cones are concerned, we may as well assume that $C = C_1$.

We consider an open subset Ω' of $\mathbb{R}^n$ such that $U_1 \cap \bar{\Omega}' = U$ and that satisfies conditions (i), (ii), and (iii) of Definition 2.1 with just one cone equal to C. Such a set clearly exists. Indeed, in view of [16, Theorem 1.2.2.2], $\partial\Omega \cap U$ is the graph of a Lipschitz function Φ from a compact subset K of a hyperplane in $\mathbb{R}^n$ into $\mathbb{R}$ (using an appropriate coordinate system (x', x_n)). If M denotes the Lipschitz constant of Φ, the standard McShane, or Whitney, extension of Φ to the whole hyperplane defined by

$$\tilde{\Phi}(x') = \min_{y \in K} (\Phi(y) + M|x' - y|)$$

provides a globally defined Lipschitz extension of Φ with the same Lipschitz constant M. It is thus sufficient to set $\Omega' = \{x \in \mathbb{R}^n; x_n < \tilde{\Phi}(x')\}$. Now, the extension of u by zero to $\Omega' \setminus \Omega$ clearly yields a function in $W^{m,p}(\Omega')$.

It follows from the previous considerations that we may assume that Ω satisfies the uniform cone condition with just one cone equal to C. Let θ_C, e_C, and h_C be, respectively, the cone's angle, outward unit axis vector, and height.

For all $0 < \varepsilon$, we now define

$$\eta(\varepsilon) = \varepsilon \sin(\theta_C/2) \tag{2.4}$$

and

$$\rho_\varepsilon(y) = \eta(\varepsilon)^{-n}\rho(y/\eta(\varepsilon)). \tag{2.5}$$

Let $\varepsilon_C = \frac{h_C}{1+\sin(\theta_C/2)}$. For all $0 < \varepsilon < \varepsilon_C$ and all $x \in \bar{\Omega}$, we then let

$$u_\varepsilon(x) = \int_{B(0,\eta(\varepsilon))} u(x - \varepsilon e - y)\rho_\varepsilon(y)\,dy. \tag{2.6}$$

By construction, we have $x - \varepsilon e - B(0,\eta(\varepsilon)) \subset x + C \subset \Omega$; therefore, $u_\varepsilon(x)$ is well defined *up to the boundary.* Moreover, since ρ_ε has support in $\bar{B}(0,\eta(\varepsilon))$, we have that

$$u_\varepsilon(x) = \int_{\mathbb{R}^n} \tilde{u}(x - \varepsilon e - y)\rho_\varepsilon(y)\,dy = \tilde{u} \star_{\varepsilon,e} \rho_\varepsilon(x),$$

i.e., $u_\varepsilon = (\tilde{u} \star_{\varepsilon,e} \rho_\varepsilon)_{|\bar{\Omega}}$, where $\tilde{u}$ is *any* extension of u to $\mathbb{R}^n$—for instance, the extension by 0. It follows that $u_\varepsilon \in C^\infty(\bar{\Omega})$. Moreover, it is easy to see that, for any multi-index α and all $x \in \bar{\Omega}$, we also have

$$\partial^\alpha u_\varepsilon(x) = \int_{\mathbb{R}^n} \widetilde{\partial^\alpha u}(x - \varepsilon e - y)\rho_\varepsilon(y)\,dy = \widetilde{\partial^\alpha u} \star_{\varepsilon,e} \rho_\varepsilon(x).$$

To conclude, it is thus sufficient to show that $u_\varepsilon \to u$ in $L^p(\Omega)$ when $\varepsilon \to 0$. This follows from the same argument that is used in standard mollification by approximating u in $L^p(\Omega)$ by a compactly supported continuous function and performing the convolution-translation on this function. □

3. A generalized version of Friedrichs's lemma.

Friedrichs's lemma was introduced to deal with partial differential equations with varying coefficients. There are many different versions of the lemma. We are concerned here with the version that estimates the commutator of the multiplication by a Lipschitz function and the convolution by a mollifier on $\mathbb{R}^n$ or on a compactly contained subset of the domain Ω; see [17] and [18]. We replace here the usual convolution by the previously introduced up-to-the-boundary convolution-translation.

In what follows, Ω is a domain in $\mathbb{R}^n$ that satisfies the uniform cone condition with just one cone, as before. For all $v \in L^p(\Omega)$, we denote by $\partial_\alpha v$ its distributional partial derivative with respect to x_α. (We thus do not use the multi-index notation.) Therefore, $\partial_\alpha v \in W^{-1,p}(\Omega)$. We need to define the convolution-translation for such distributions:

$$\partial_\alpha v \star_{\varepsilon,e} \rho_\varepsilon(x) = \int_{B(0,\eta(\varepsilon))} v(x - \varepsilon e - y)\partial_\alpha \rho_\varepsilon(y)\,dy, \tag{3.1}$$

where $\eta(\varepsilon)$ and ρ_ε are defined as in formulas (2.4) and (2.5). Clearly, this definition agrees with (2.6) when v is C^∞. Moreover, the resulting function is in $C^\infty(\bar\Omega)$. Finally, if we take a sequence $v_k \in C^\infty(\bar\Omega)$ such that $v_k \to v$ in $L^p(\Omega)$ when $k \to +\infty$, then $\partial_\alpha v_k \star_{\varepsilon,e} \rho_\varepsilon \to \partial_\alpha v \star_{\varepsilon,e} \rho_\varepsilon$ in $L^p(\Omega)$.

Our version of Friedrichs's lemma is as follows.

Lemma 3.1. *Let $v \in L^p(\Omega)$ and $a \in W^{1,\infty}(\Omega)$; then there exists a constant M which depends only on ρ and on the cone angle such that*

$$\|(a\partial_\alpha v) \star_{\varepsilon,e} \rho_\varepsilon - a(\partial_\alpha v \star_{\varepsilon,e} \rho_\varepsilon)\|_{L^p(\Omega)} \le M\|a\|_{W^{1,\infty}(\Omega)}\|v\|_{L^p(\Omega)}. \tag{3.2}$$

Proof. Note first that if $v \in L^p(\Omega)$ and $a \in W^{1,\infty}(\Omega)$, then $a\partial_\alpha v \in W^{-1,p}(\Omega)$ so that all terms are well defined. Moreover, in view of the above remarks, it is sufficient to establish estimate (3.2) for functions v in $\mathcal{D}(\Omega)$, which is a dense subset of $L^p(\Omega)$. In this case, $a\partial_\alpha v$ belongs to $W^{1,\infty}(\Omega)$ with compact support.

Let us estimate the commutator. For all $x \in \bar\Omega$, we have

$$[(a\partial_\alpha v)\star_{\varepsilon,e}\rho_\varepsilon - a(\partial_\alpha v\star_{\varepsilon,e}\rho_\varepsilon)](x) = \int_{B(0,\eta(\varepsilon))}[a(x-\varepsilon e-y)-a(x)]\partial_\alpha v(x-\varepsilon e-y)\rho_\varepsilon(y)\,dy. \tag{3.3}$$

Integrating the right-hand side of (3.3) by parts in the ball, we obtain

$$\begin{aligned}[(a\partial_\alpha v) \star_{\varepsilon,e} \rho_\varepsilon - a(\partial_\alpha v \star_{\varepsilon,e} \rho_\varepsilon)](x) &= \int_{B(0,\eta(\varepsilon))} \partial_\alpha a(x-\varepsilon e-y)v(x-\varepsilon e-y)\rho_\varepsilon(y)\,dy\\ &\quad+\int_{B(0,\eta(\varepsilon))}[a(x-\varepsilon e-y)-a(x)]v(x-\varepsilon e-y)\partial_\alpha\rho_\varepsilon(y)\,dy.\end{aligned} \tag{3.4}$$

Note that both integral quantities vanish for x outside of a compact neighborhood K of the support of v that is independent of ε for $\varepsilon \le \varepsilon_C$. We denote by $\tilde{\tilde a}$ the McShane extension of the restriction of a to K, and by $\tilde v$ the extension of v by zero. It is then clear that for all $x \in \bar\Omega$,

$$[(a\partial_\alpha v) \star_{\varepsilon,e} \rho_\varepsilon - a(\partial_\alpha v \star_{\varepsilon,e} \rho_\varepsilon)](x) = f_1(x) + f_2(x), \tag{3.5}$$

where

$$f_1(x) = \int_{\mathbb{R}^n} \partial_\alpha\tilde{\tilde a}(x-\varepsilon e-y)\tilde v(x-\varepsilon e-y)\rho_\varepsilon(y)\,dy$$

and

$$f_2(x) = \int_{\mathbb{R}^n}[\tilde{\tilde a}(x-\varepsilon e-y)-\tilde{\tilde a}(x)]\tilde v(x-\varepsilon e-y)\partial_\alpha\rho_\varepsilon(y)\,dy$$

for all $x \in \mathbb{R}^n$. It is thus enough to estimate f_1 and f_2 in L^p separately.

For the first term, we have

$$\begin{aligned}|f_1(x)| &\le \int_{\mathbb{R}^n}|\partial_\alpha\tilde{\tilde a}(x-\varepsilon e-y)||\tilde v(x-\varepsilon e-y)|\rho_\varepsilon(y)\,dy\\ &\le \|\partial_\alpha\tilde{\tilde a}\|_{L^\infty(\mathbb{R}^n)}\int_{\mathbb{R}^n}|\tilde v(x-\varepsilon e-y)|\rho_\varepsilon(y)\,dy\\ &\le \|a\|_{W^{1,\infty}(\Omega)}(|\tilde v|\star\rho_\varepsilon)(x-\varepsilon e)\end{aligned}$$

for all $x \in \mathbb{R}^n$. Therefore,

$$\begin{aligned}\|f_1\|_{L^p(\Omega)} &\leq \|f_1\|_{L^p(\mathbb{R}^n)}\\ &\leq \|a\|_{W^{1,\infty}(\Omega)}\|(|\tilde{v}| \star \rho_\varepsilon)\|_{L^p(\mathbb{R}^n)}\\ &\leq \|a\|_{W^{1,\infty}(\Omega)}\|\tilde{v}\|_{L^p(\mathbb{R}^n)}\\ &= \|a\|_{W^{1,\infty}(\Omega)}\|v\|_{L^p(\Omega)}.\end{aligned}$$

Similarly, for the second term,

$$\begin{aligned}|f_2(x)| &\leq \int_{\mathbb{R}^n} |\tilde{\tilde{a}}(x-\varepsilon e-y) - \tilde{\tilde{a}}(x)|\,|\tilde{v}(x-\varepsilon e-y)|\,|\partial_\alpha\rho_\varepsilon(y)|\,dy\\ &\leq \|\nabla\tilde{\tilde{a}}\|_{L^\infty(\mathbb{R}^n)}\int_{\mathbb{R}^n} |\tilde{v}(x-\varepsilon e-y)|\,|\varepsilon e - y|\,|\partial_\alpha\rho_\varepsilon(y)|\,dy\\ &\leq \|a\|_{W^{1,\infty}(\Omega)}(|\tilde{v}| \star g_\varepsilon)(x-\varepsilon e)\end{aligned}$$

for all $x \in \mathbb{R}^n$, where

$$g_\varepsilon(y) = |\varepsilon e - y|\,|\partial_\alpha\rho_\varepsilon(y)|.$$

In view of definitions (2.4) and (2.5), it is easy to see that

$$\|g_\varepsilon\|_{L^1(\mathbb{R}^n)} \leq \left(\frac{1}{\sin(\theta_C/2)} + 1\right)\|\partial_\alpha\rho\|_{L^1(\mathbb{R}^n)}.$$

Therefore,

$$\|f_2\|_{L^p(\Omega)} \leq \left(\frac{1}{\sin(\theta_C/2)} + 1\right)\|\partial_\alpha\rho\|_{L^1(\mathbb{R}^n)}\|a\|_{W^{1,\infty}(\Omega)}\|v\|_{L^p(\Omega)},$$

hence the result with $M = 1 + \left(\frac{1}{\sin(\theta_C/2)} + 1\right)\|\partial_\alpha\rho\|_{L^1(\mathbb{R}^n)}$. □

The following corollary will be the basic tool for our density results in the context of the Koiter shell model.

COROLLARY 3.2. *For all $v \in L^p(\Omega)$ and $a \in W^{1,\infty}(\Omega)$,*

$$\|(a\partial_\alpha v) \star_{\varepsilon,e} \rho_\varepsilon - a(\partial_\alpha v \star_{\varepsilon,e} \rho_\varepsilon)\|_{L^p(\Omega)} \to 0 \quad \textit{when} \quad \varepsilon \to 0. \tag{3.6}$$

Proof. Proceed as in [18] by approximating v in $L^p(\Omega)$ by a sequence of functions in $\mathcal{D}(\Omega)$. □

4. Application to the Koiter shell model.

4.1. Formulation of the problem. In this section, we briefly recall the formulation of the linear Koiter shell model introduced in [5] and [6]. This formulation is much simpler than the classical formulation and is, furthermore, valid for midsurfaces that can have discontinuous curvatures. We refer to [3] and [8] for general elastic shell theory.

In what follows, Greek indices and exponents always belong to the set $\{1, 2\}$, while Latin indices and exponents belong to the set $\{1, 2, 3\}$. We use the Einstein summation convention unless otherwise specified.

Let ω denote a Lipschitz domain of $\mathbb{R}^2$. We consider a shell with midsurface $S = \varphi(\bar{\omega})$, where $\varphi \in W^{2,\infty}(\omega; \mathbb{R}^3)$ is a one-to-one mapping such that the two vectors $a_\alpha = \partial_\alpha\varphi$ are linearly independent at each point $x \in \bar{\omega}$. We let $a_3 = a_1 \wedge a_2/|a_1 \wedge a_2|$ be the unit normal vector on the midsurface at point $\varphi(x)$. The vectors a_i define the covariant basis at point $\varphi(x)$. The regularity of the midsurface chart and the hypothesis of linear independence on $\bar{\omega}$ imply that the vectors a_i belong to $W^{1,\infty}(\omega;\mathbb{R}^3)$. The contravariant basis a^i is defined by the relations

$$a^i(x)\cdot a_j(x)=\delta^i_j,$$

where δ^i_j is the Kronecker symbol. In particular, $a^3(x)=a_3(x)$. As before, $a^i\in W^{1,\infty}(\omega;\mathbb{R}^3)$. We let $a(x)=|a_1(x)\wedge a_2(x)|^2$ so that $\sqrt{a}$ is the area element of the midsurface in the chart φ.

The first fundamental form of the surface is given in covariant components by $a_{\alpha\beta}=a_\alpha\cdot a_\beta\in W^{1,\infty}(\omega)$. The Christoffel symbols of the midsurface are given by $\Gamma^\rho_{\alpha\beta}=\Gamma^\rho_{\beta\alpha}=a^\rho\cdot\partial_\beta a_\alpha$, and we have $\Gamma^\rho_{\alpha\beta}\in L^\infty(\omega)$.

Let us recall the new expressions for the various strain tensors that were introduced in [5] and [6]. Let $u\in H^1(\omega;\mathbb{R}^3)$ be a displacement of the midsurface, i.e., a regular mapping from $\bar{\omega}$ into $\mathbb{R}^3$. Its linearized strain tensor is given by $\gamma(u)=\gamma_{\alpha\beta}(u)a^\alpha\otimes a^\beta$ with

$$\gamma_{\alpha\beta}(u)=\frac{1}{2}(\partial_\alpha u\cdot a_\beta+\partial_\beta u\cdot a_\alpha)\in L^2(\omega), \tag{4.1}$$

and its linearized change of curvature tensor is given by $\Upsilon(u)=\Upsilon_{\alpha\beta}(u)a^\alpha\otimes a^\beta$ with

$$\Upsilon_{\alpha\beta}(u)=(\partial_{\alpha\beta}u-\Gamma^\rho_{\alpha\beta}\partial_\rho u)\cdot a_3\in H^{-1}(\omega). \tag{4.2}$$

See [6] for a comparison with the classical approach, in which the displacement is identified with the triple of its covariant components, and an explanation of why our new approach does not require φ to be of class at least C^3, which includes many interesting cases, as we mentioned earlier.

In [5] and [6], we introduced the function space

$$W=\Big\{v\in H^1(\omega;\mathbb{R}^3),\partial_{\alpha\beta}v\cdot a_3\in L^2(\omega)\Big\} \tag{4.3}$$

for shell displacements. Note that if $v\in H^1(\omega;\mathbb{R}^3)$, then $\partial_{\alpha\beta}v\cdot a_3$ is a priori in $H^{-1}(\omega)$. In view of formulas (4.1) and (4.2), it is apparent that displacements in W are such that their linearized strain and change of curvature tensors are square-integrable. When equipped with its natural norm

$$\|v\|_W=\left(\|v\|^2_{H^1(\omega;\mathbb{R}^3)}+\sum_{\alpha,\beta}\|\partial_{\alpha\beta}v\cdot a_3\|^2_{L^2(\omega)}\right)^{1/2}, \tag{4.4}$$

the space W is a Hilbert space. To formulate an equilibrium problem for the shell, we consider $e>0$ to be the thickness of the shell and an elasticity tensor $a^{\alpha\beta\rho\sigma}\in L^\infty(\omega)$, which we assume to satisfy the usual symmetries and to be uniformly strictly positive.

In terms of boundary conditions, the simplest case is that of a shell clamped on all of its boundaries. This corresponds to the space

$$V_1=\big\{v\in W;v=\partial_\alpha v\cdot a_3=0\text{ on }\partial\omega\big\}, \tag{4.5}$$

which is a closed subspace of W, endowed with the norm of W. In [6], we proved the following existence and uniqueness result for Koiter's model.

THEOREM 4.1. *Let $f \in L^2(\omega;\mathbb{R}^3)$ be a given force resultant density. Then there exists a unique solution to the variational problem: Find $u \in V_1$ such that*

$$\forall v \in V_1, \quad \int_\omega e a^{\alpha\beta\rho\sigma}\left(\gamma_{\alpha\beta}(u)\gamma_{\rho\sigma}(v) + \frac{e^2}{12}\Upsilon_{\alpha\beta}(u)\Upsilon_{\rho\sigma}(v)\right)\sqrt{a}\,dx = \int_\omega f\cdot v\sqrt{a}\,dx. \tag{4.6}$$

Remark 4.2. 1. The proof of this theorem relies on the new version of the rigid displacement lemma and the Korn inequality for surfaces with $W^{2,\infty}$ regularity; see [6] for details.

2. Also in [6], it is proved that the space W defines an extension of the classical framework of [4] to our case. Indeed, when $\varphi \in C^3$, the function space of [4] is canonically isomophic to the space V_1. Moreover, the new and classical expressions for the linearized strain and change of curvature tensors coincide under this isomorphism. Consequently, the solution given by Theorem 4.1 is in this case equal, modulo the isomorphism, to the solution found in [4].

3. The case of a shell clamped on a part of its boundary $\gamma_0 \subset \partial\omega$ and submitted to tractions and moments on the remaining part is also treated in [6]. The relevant function space is

$$V_{1,\gamma_0} = \{v \in W; v = \partial_\alpha v \cdot a_3 = 0 \text{ on } \gamma_0\}, \tag{4.7}$$

which is also a closed subspace of W. For a simply supported shell, the relevant space is simply

$$V_0 = \{v \in W; v = 0 \text{ on } \partial\omega\}. \tag{4.8}$$

We thus also obtain existence and uniqueness results in these cases. □

4.2. Density results. One fundamental issue that was not addressed in [6] is the density of smooth functions in the various function spaces introduced in our new formulation of the Koiter model. The density of smooth functions is, for instance, required in order to make sure that standard finite element methods will actually approximate the solution of the continuous problem. Another use of this density will be to show the continuous dependence of the solution of the model on the midsurface, in an appropriate sense. In [6], the consistency of our formulation with the classical formulation is an a priori consistency: we know that our formulation is more general than and coincides with the classical formulation when both are applicable. Continuous dependence is a way to prove a posteriori consistency via a convergence result.

It should be noted that such a density result cannot be taken for granted since these spaces are not of a standard kind. Similar questions arise in transport theory; see [2], [12], and [15], for example. In the case of the transport equation, the definition of the relevant function spaces involves a directional derivative of the form $a\nabla u$, with a a vector field and u a scalar unknown, that is required to satisfy some integrability condition. Although formally slightly reminiscent of this situation, our function space setting is different, since the quantities of interest in shell theory, $\partial_{\alpha\beta}u \cdot a_3$, are not directional derivatives—a_3 does not "live" in the same space as u—and we cannot adapt techniques based on this special structure to our case.

It should also be noted that if $\omega = \mathbb{R}^2$, then the density of smooth functions in W follows more or less readily from the classical version of Friedrichs's lemma (see [18]), as will be made clear in the ensuing proofs. However, from the point of view of the applications, a shell whose midsurface was described by a chart over $\mathbb{R}^2$ would be of little interest. Prescribing tractions and moments on the boundary would be difficult, and the shell would be diffeomorphic to an open disk, which is restrictive in terms of the topology of the shell since multiply connected shells would not be allowed. Thus there is no escaping the difficulties that arise at the boundary.

As we mentioned earlier, the standard way of performing mollification up to the boundary consists in using an extension operator and then mollifying over $\mathbb{R}^2$. This does not seem to be of much help here. Indeed, if $E_1(u)$ and $E_2(\varphi)$ denote, respectively, an H^1-extension operator for the displacement u and a $W^{2,\infty}$-extension operator for the chart φ to $\mathbb{R}^2$, there does not seem to be an easy way of devising E_1 and E_2 in a way that ensures that $\partial_{\alpha\beta}(E_1(u)) \cdot \tilde{a}_3$ will belong to $L^2(\mathbb{R}^2)$, where $\tilde{a}_3$ denotes the corresponding extended normal vector (assuming it is defined), whenever $\partial_{\alpha\beta} u \cdot a_3$ belongs to $L^2(\omega)$. The classical techniques using reflections or integral operators do not seem to work very well because of the product of two quantities. The same remark applies if we try to extend a_3 itself without any reference to the geometrical underpinnings of the situation.

It is this failure that prompted us to look for an alternative and eventually rediscover a mollification technique essentially already put forth in [23].

Let us start with the larger function space, without boundary conditions.

THEOREM 4.3. *Assume that ω satisfies the uniform cone condition. Then the space $C^\infty(\bar{\omega};\mathbb{R}^3)$ is dense in W.*

Proof. First, it is clear that $C^\infty(\bar{\omega};\mathbb{R}^3) \subset W$. Let $u \in W$. We want to construct a sequence u_ε of $C^\infty(\bar{\omega};\mathbb{R}^3)$-functions that converges to u in the norm of W, i.e., such that $u_\varepsilon \to u$ in $H^1(\omega;\mathbb{R}^3)$ and $\partial_{\alpha\beta} u_\varepsilon \cdot a_3 \to \partial_{\alpha\beta} u \cdot a_3$ in $L^2(\omega)$ for all indices α, β.

It is not difficult to check that the space W can be localized using a partition of unity that is adapted to the uniform cone condition satisfied by ω. We can thus assume that u is compactly supported in one of the sets $U_i \cap \bar{\omega}$ introduced in the proof of Theorem 2.4. We leave the case of the "interior" part in $U_0 \cap \bar{\omega}$ aside for the time being.

Let $U = U_1 \cap \bar{\omega}$. As in the proof of Theorem 2.4, we can assume that ω satisfies the uniform cone condition with just one cone and that u is compactly supported in U. Then introducing

$$u_\varepsilon = \tilde{u} \star_{\varepsilon,e} \rho_\varepsilon,$$

Theorem 2.4 shows that

$$u_\varepsilon \to u \quad \text{in} \quad H^1(\omega;\mathbb{R}^3).$$

Let u_i, $u_{\varepsilon,i}$, and $a_{3,i}$ denote the Cartesian components of u, u_ε, and a_3, respectively, so that $\partial_{\alpha\beta} u \cdot a_3 = (\partial_{\alpha\beta} u_i) a_{3,i}$. Applying Corollary 3.2 to $\partial_\beta u_i \in L^2(\omega)$, we obtain

$$\|(a_{3,i}\partial_{\alpha\beta} u_i) \star_{\varepsilon,e} \rho_\varepsilon - a_{3,i}((\partial_{\alpha\beta} u_i) \star_{\varepsilon,e} \rho_\varepsilon)\|_{L^2(\omega)} \longrightarrow 0 \text{ when } \varepsilon \to 0.$$

Now, since $u \in W$, we also have that $(a_{3,i}\partial_{\alpha\beta} u_i) \in L^2(\omega)$. Therefore, by Theorem 2.4, it follows that

$$\|a_{3,i}\partial_{\alpha\beta} u_i - (a_{3,i}\partial_{\alpha\beta} u_i) \star_{\varepsilon,e} \rho_\varepsilon\|_{L^2(\omega)} \longrightarrow 0 \text{ when } \varepsilon \to 0$$

as well. Since

$$\partial_{\alpha\beta} u_{i,\varepsilon} = (\partial_{\alpha\beta} u_i) \star_{\varepsilon,e} \rho_\varepsilon,$$

we have thus shown that

$$\|a_{3,i}\partial_{\alpha\beta} u_i - a_{3,i}\partial_{\alpha\beta} u_{i,\varepsilon}\|_{L^2(\omega)} \longrightarrow 0 \text{ when } \varepsilon \to 0,$$

which concludes the proof near the boundary.

Concerning the interior part of u, it is apparent that the same proof works using standard mollification and the classical Friedrichs lemma. □

Let us now consider the case of various boundary conditions. As before, we assume that ω satisfies the uniform cone condition.

THEOREM 4.4. *For a totally clamped shell, $\mathcal{D}(\omega;\mathbb{R}^3)$ is dense in V_1.*

Proof. Localize as before near the boundary. We claim that the extension $\tilde{u}$ of u by zero to the whole of $\mathbb{R}^2$ is such that $\tilde{u} \in H^1(\mathbb{R}^2;\mathbb{R}^3)$ and $\partial_{\alpha\beta}\tilde{u}\cdot\tilde{\tilde{a}}_3 \in L^2(\mathbb{R}^2)$. Indeed, both conditions are equivalent to having $\tilde{u} \in H^1(\mathbb{R}^2;\mathbb{R}^3)$ and $\partial_\alpha\tilde{u}\cdot\tilde{\tilde{a}}_3 \in H^1(\mathbb{R}^2)$. Since both functions are piecewise H^1 and have no jump on $\partial\omega$, the claim is true.

Instead of translating inside the domain as earlier, we translate here *outside* and let $u_\varepsilon = \tilde{u} \star_{\varepsilon,-e} \rho_\varepsilon$. The same proof as in Theorem 4.3 shows that $(u_\varepsilon)_{|\omega} \to u$ in W and that $u_\varepsilon \in C^\infty(\mathbb{R}^2;\mathbb{R}^3)$. Moreover, since $\tilde{u}$ is identically zero outside of ω, it is clear that u_ε has compact support in ω. It may be necessary to change the cone in such a way that the exterior cone condition is also satisfied to achieve this, which is possible since ω is locally Lipschitz. Hence the result. □

Theorem 4.4 above can be seen as an intermediary step for the following density result.

THEOREM 4.5. *Assume that γ_0 consists of a finite union of open arcs in $\partial\omega$, and let $C^\infty_{c,\gamma_0}(\bar{\omega};\mathbb{R}^3)$ denote the set of functions in $C^\infty(\bar{\omega};\mathbb{R}^3)$ that are equal to 0 in a neighborhood of γ_0. Then $C^\infty_{c,\gamma_0}(\bar{\omega};\mathbb{R}^3)$ is dense in V_{1,γ_0}.*

Proof. We localize as before around γ_0, the interior of its complement in $\partial\omega$, and the endpoints of γ_0. Clearly, for the parts localized around γ_0, the same argument as in the proof of Theorem 4.4 applies. Equally clearly, for the parts localized around the interior of the complement, the argument of the proof of Theorem 4.3 applies. What remains are the parts that are localized around the endpoints of γ_0.

Let us thus assume that 0 is such an endpoint, and let us localize u in a ball of radius ε around this point. To this end, we introduce a function $\theta \in \mathcal{D}(B(0,1))$ such that $\theta(x) = 1$ if $|x| \leq 1/2$ and $\theta(x) = 0$ for $|x| \geq 3/4$, and let $\theta_\varepsilon(x) = \theta(x/\varepsilon)$. We want to show that $\theta_\varepsilon u$ tends to zero strongly in W when $\varepsilon \to 0$, so that we can approximate u by 0 in $B(0,\varepsilon/2)$.

Since u and $\partial_\alpha u \cdot a_3$ vanish on an arc that has 0 as an endpoint, we can apply Poincaré's inequality to both quantities to deduce that

$$\|u\|^2_{L^2(B(0,\varepsilon);\mathbb{R}^3)} \leq \varepsilon^2 \|\nabla u\|^2_{L^2(B(0,\varepsilon);M_{32})} \tag{4.9}$$

and

$$\|\partial_\alpha u \cdot a_3\|^2_{L^2(B(0,\varepsilon))} \leq \varepsilon^2 \|\nabla(\partial_\alpha u \cdot a_3)\|^2_{L^2(B(0,\varepsilon);\mathbb{R}^2)}. \tag{4.10}$$

By estimate (4.9), we see that $\theta_\varepsilon u \to 0$ in $H^1(\omega;\mathbb{R}^3)$. Indeed, $\partial_\alpha(\theta_\varepsilon u) = (\partial_\alpha\theta_\varepsilon)u + \theta_\varepsilon\partial_\alpha u$, and

$$\|(\partial_\alpha\theta_\varepsilon)u\|^2_{L^2(B(0,\varepsilon);\mathbb{R}^3)} \leq \|\nabla u\|^2_{L^2(B(0,\varepsilon);M_{32})} \to 0 \text{ when } \varepsilon \to 0.$$

We now note that $u \cdot a_3$ also vanishes on the same arc so that by Poincaré's inequality

$$\|u \cdot a_3\|^2_{L^2(B(0,\varepsilon))} \leq \varepsilon^2 \|\nabla(u \cdot a_3)\|^2_{L^2(B(0,\varepsilon);\mathbb{R}^2)}. \tag{4.11}$$

Now $\partial_\alpha(u \cdot a_3) = \partial_\alpha u \cdot a_3 + u \cdot \partial_\alpha a_3$, and therefore

$$\|\partial_\alpha(u\cdot a_3)\|^2_{L^2(B(0,\varepsilon))} \leq 2\big(\|\partial_\alpha u\cdot a_3\|^2_{L^2(B(0,\varepsilon))}+\|u\cdot\partial_\alpha a_3\|^2_{L^2(B(0,\varepsilon))}\big) \leq C\varepsilon^2\|u\|^2_{W(B(0,\varepsilon))},$$

using estimate (4.10) for the first term and estimate (4.9) and the fact that $\partial_\alpha a_3 \in L^\infty(\omega)$ for the second term, where $\|\cdot\|_{W(B(0,\varepsilon))}$ denotes the local W-norm on $B(0,\varepsilon)$. Consequently, by estimate (4.11), we obtain

$$\|u \cdot a_3\|^2_{L^2(B(0,\varepsilon))} \leq C\varepsilon^4 \|u\|^2_{W(B(0,\varepsilon))}. \tag{4.12}$$

We have

$$\partial_{\alpha\beta}(\theta_\varepsilon u) \cdot a_3 = (\partial_{\alpha\beta}\theta_\varepsilon) u \cdot a_3 + (\partial_\alpha \theta_\varepsilon)(\partial_\beta u) \cdot a_3 + (\partial_\beta\theta_\varepsilon)(\partial_\alpha u) \cdot a_3 + \theta_\varepsilon(\partial_{\alpha\beta} u) \cdot a_3.$$

Therefore, putting estimates (4.10) and (4.12) together, we see that

$$\|\partial_{\alpha\beta}(\theta_\varepsilon u) \cdot a_3\|_{L^2(B(0,\varepsilon))} \leq C\|u\|_{W(B(0,\varepsilon))} \to 0 \text{ when } \varepsilon \to 0,$$

which shows that $\theta_\varepsilon u \to 0$ in W.

Finally, it is fairly clear that the elements of the sequence u_ε, which are reconstructed by patching together all the local approximations of u, belong to $C^\infty_{c,\gamma_0}(\bar{\omega};\mathbb{R}^3)$, and the theorem is proved. □

Remark 4.6. Note that the space $C^\infty_{c,\gamma_0}(\bar{\omega};\mathbb{R}^3)$ does not depend on the chart φ, whereas the space V_{1,γ_0} does. This is useful since one of the applications we have in mind is the dependence of the solution of Koiter's model on the midsurface. The space $C^\infty_{c,\gamma_0}(\bar{\omega};\mathbb{R}^3)$ is a common dense subspace of all possible V_{1,γ_0} spaces for all possible midsurfaces. □

The case of a simply supported shell actually seems to be more difficult. We only solve it here for a domain ω of class C^∞ and by resorting to classical techniques.

THEOREM 4.7. *Assume that ω is of class C^∞. Then $C^\infty(\bar{\omega};\mathbb{R}^3) \cap H^1_0(\omega;\mathbb{R}^3)$ is dense in V_0.*

Proof. We proceed in a classical fashion. First localize as before. For the parts near the boundary, we can thus assume that $\omega = \{(x_1,x_2) \in \mathbb{R}^2; x_2 < \Psi(x_1)\}$, where $\Psi\colon \mathbb{R} \to \mathbb{R}$ is of class C^∞. Next we flatten the boundary using the C^∞-diffeomorphism

$$\begin{cases} \Theta_1(x) = x_1, \\ \Theta_2(x) = x_2 - \Psi(x_1). \end{cases}$$

This obviously induces an isomorphism on the associated V_0 spaces so that we are reduced to the case $\omega = \mathbb{R} \times \mathbb{R}^*_-$. We now extend u and a_3 for $x_2 > 0$ by

$$\begin{cases} \tilde{u}(x_1,x_2) = -u(x_1,-x_2), \\ \tilde{a}_3(x_1,x_2) = a_3(x_1,-x_2), \end{cases}$$

respectively. Clearly, $\tilde{u} \in H^1(\mathbb{R}^2;\mathbb{R}^3)$, $\tilde{a}_3 \in W^{1,\infty}(\mathbb{R}^2;\mathbb{R}^3)$, and $\tilde{u}$ is odd with respect to x_2.

Let us show that $\partial_{\alpha\beta}\tilde{u}\cdot\tilde{a}_3$ belongs to $L^2(\mathbb{R}^2)$, or, equivalently, that $\partial_\alpha\tilde{u}\cdot\tilde{a}_3$ is in $H^1(\mathbb{R}^2)$. For $x_2>0$, we have

$$\begin{cases}\partial_1\tilde{u}\cdot\tilde{a}_3(x_1,x_2)=-\partial_1 u\cdot a_3(x_1,-x_2),\\ \partial_2\tilde{u}\cdot\tilde{a}_3(x_1,x_2)=\partial_2 u\cdot a_3(x_1,-x_2),\end{cases}$$

so that $(\partial_\alpha\tilde{u}\cdot\tilde{a}_3)_{|\mathbb{R}\times\mathbb{R}_+^*}$ belongs to $H^1(\mathbb{R}\times\mathbb{R}_+^*)$. It thus suffices to prove that the jump of both quantities across $x_2=0$ is zero. This is clear for the second one as the traces at $x_2=0$ of both sides of the equality obviously coincide. It can also be shown with a little more work that $\partial_1 u\cdot a_3$ considered as a function in the variable x_2 belongs to $C^0(\mathbb{R}_-;H^{-1}(\mathbb{R}))$ with $\partial_1 u\cdot a_3(0)=0$, and thus there is no jump either.

We can now introduce a radial mollifier ρ, define $\rho_\varepsilon=\varepsilon^{-2}\rho(\cdot/\varepsilon)$, and let $u_\varepsilon=\tilde{u}\star\rho_\varepsilon$. Clearly, $u_{\varepsilon|\omega}\in C^\infty(\bar{\omega};\mathbb{R}^3)$, $u_{\varepsilon|\omega}\to u$ in W as $\varepsilon\to 0$ by the classical Friedrichs lemma, and $u_\varepsilon(x_1,0)=0$ by the imparity of u and parity of ρ with respect to x_2.

We can go back to the original domain by composition with the C^∞-diffeomorphism Θ^{-1}. □

Remark 4.8. 1. Since we are mainly interested in finding a dense subspace that does not depend on the midsurface, the above proof shows that if ω is of class $W^{2,\infty}$, then $W^{2,\infty}(\omega;\mathbb{R}^3)\cap H_0^1(\omega;\mathbb{R}^3)$ is dense in V_0.

2. The above proof also works for piecewise C^∞ domains satisfying the uniform cone condition, for instance polygons, by performing adequate reflections at the angles. In this case, $C^\infty(\bar{\omega};\mathbb{R}^3)\cap H_0^1(\omega;\mathbb{R}^3)$ is also dense in V_0. □

4.3. Continuous dependence on the midsurface. Our goal in this section is to show that the formulation of Koiter's model we proposed in [5] and [6] provides an adequate extension of the classical formulation for C^3-midsurfaces to $W^{2,\infty}$-midsurfaces, at least from the mathematical point of view. The density results of the previous section were formulated for a domain that satisfies the uniform cone condition. For practical purposes in the case of shells, we will from now on consider only bounded Lipschitz domains.

Let us thus consider a sequence of midsurface charts φ^n that approximate a given chart φ in the sense that $\varphi^n\to\varphi$ in $W^{2,p}(\omega;\mathbb{R}^3)$ strong for all $1<p<+\infty$ and $\varphi^n\overset{*}{\rightharpoonup}\varphi$ in $W^{2,\infty}(\omega;\mathbb{R}^3)$ weak-$*$. (Note that for any $\varphi\in W^{2,\infty}(\omega;\mathbb{R}^3)$, it is possible to construct such a sequence with φ^n of class C^3.) All corresponding geometric quantities will from now on be indicated by an n superscript. For instance, the covariant basis vectors are denoted by a_i^n, the covariant components of the first fundamental form and the area element by $a_{\alpha\beta}^n$ and $\sqrt{a^n}$, respectively, and the Christoffel symbols by $\Gamma_{\alpha\beta}^{n,\rho}$. We assume for simplicity that all shells have the same thickness and the same Lamé moduli μ and λ and denote by $a^{n,\alpha\beta\rho\sigma}$ the contravariant components of the elasticity tensor

$$a^{n,\alpha\beta\rho\sigma}=2\mu(a^{n,\alpha\beta}a^{n,\rho\sigma}+a^{n,\alpha\sigma}a^{n,\beta\sigma})+\frac{4\lambda\mu}{\lambda+2\mu}a^{n,\alpha\beta}a^{n,\rho\sigma}, \tag{4.13}$$

where $a^{n,\alpha\beta}$ denote the contravariant components of the first fundamental form. Finally, for all displacements v of the shells, we denote by

$$\gamma_{\alpha\beta}^n(v)=\frac{1}{2}(\partial_\alpha v\cdot a_\beta^n+\partial_\beta v\cdot a_\alpha^n)$$

and

$$\Upsilon_{\alpha\beta}^n(v)=(\partial_{\alpha\beta}v-\Gamma_{\alpha\beta}^{n,\rho}\partial_\rho v)\cdot a_3^n$$

the covariant components of the strain and change of curvature tensors with explicit dependence on the charts.

Let us collect all the information on convergence properties of the various geometric and mechanical quantities that we will need later on in one lemma.

LEMMA 4.9. *Let φ^n be a sequence of charts such that $\varphi^n \to \varphi$ in $W^{2,p}(\omega;\mathbb{R}^3)$ strong for all $1 < p < +\infty$ and $\varphi^n \overset{*}{\rightharpoonup} \varphi$ in $W^{2,\infty}(\omega;\mathbb{R}^3)$ weak-$*$. Then*

$$a_i^n \to a_i \textit{ strongly in } W^{1,p}(\omega;\mathbb{R}^3)\ \forall\ 1 < p < +\infty \textit{ and weakly-}* \textit{ in } W^{1,\infty}(\omega;\mathbb{R}^3), \tag{4.14}$$

$$a_{\alpha\beta}^n \to a_{\alpha\beta}, \sqrt{a^n} \to \sqrt{a}, \textit{ and } a^{n,\alpha\beta\rho\sigma} \to a^{\alpha\beta\rho\sigma} \textit{ in } C^0(\bar{\omega}), \tag{4.15}$$

and

$$\Gamma_{\alpha\beta}^{n,\rho} \to \Gamma_{\alpha\beta}^{\rho} \textit{ strongly in } L^p(\omega)\ \forall\ 1 < p < +\infty \textit{ and weakly-}* \textit{ in } L^\infty(\omega). \tag{4.16}$$

Proof. The proof is clear, using Morrey's theorem and the fact that the covariant tangent vectors are assumed to be linearly independent. □

In what follows, we will concentrate on the case of a totally clamped shell submitted only to force resultants for brevity. All results remain true—with appropriate modifications—for a simply supported shell and a partially clamped shell submitted to edge tractions and moments on the free part of the boundary. Let us rewrite the spaces involved with explicit dependence on the charts. For all n, we thus let $W^n = \{v \in H^1(\omega;\mathbb{R}^3), \partial_{\alpha\beta} v \cdot a_3^n \in L^2(\omega)\}$, equipped with their natural norm $\|v\|_{W^n} = \big(\|v\|^2_{H^1(\omega;\mathbb{R}^3)} + \sum_{\alpha,\beta} \|\partial_{\alpha\beta} v \cdot a_3^n\|^2_{L^2(\omega)}\big)^{1/2}$, and $V_1^n = \{v \in W^n; v = \partial_\alpha v \cdot a_3^n = 0 \text{ on } \partial\omega\}$, which is a closed subspace of W^n for all n.

For $f^n \in L^2(\omega;\mathbb{R}^3)$ we let u^n be the unique solution to the variational formulation of Koiter's model: Find $u^n \in V_1^n$ such that

$$\forall v^n \in V_1^n,\ \int_\omega e a^{n,\alpha\beta\rho\sigma} \left(\gamma_{\alpha\beta}^n(u^n)\gamma_{\rho\sigma}^n(v^n) + \frac{e^2}{12}\Upsilon_{\alpha\beta}^n(u^n)\Upsilon_{\rho\sigma}^n(v^n) \right) \sqrt{a^n}\, dx$$
$$= \int_\omega f^n \cdot v^n \sqrt{a^n}\, dx. \tag{4.17}$$

Our main result is the following.

THEOREM 4.10. *Let φ^n be a sequence of charts such that $\varphi^n \to \varphi$ in $W^{2,p}(\omega;\mathbb{R}^3)$ strong for all $1 < p < +\infty$ and $\varphi^n \overset{*}{\rightharpoonup} \varphi$ in $W^{2,\infty}(\omega;\mathbb{R}^3)$ weak-$*$, and let f^n be a sequence of force resultant densities such that $f^n \to f$ in $L^2(\omega;\mathbb{R}^3)$. Then*

$$u^n \to u \textit{ in } H^1(\omega;\mathbb{R}^3) \quad \textit{and} \quad \Upsilon_{\alpha\beta}^n(u^n) \to \Upsilon_{\alpha\beta}(u) \textit{ in } L^2(\omega), \tag{4.18}$$

where u is the solution to Koiter's model for a clamped shell with midsurface chart φ and applied force resultant density f.

The proof is comprised of a series of lemmas.

LEMMA 4.11. *If $v^n \in H^1(\omega;\mathbb{R}^3)$ and $\varphi^n \in W^{2,\infty}(\omega;\mathbb{R}^3)$ are two sequences such that $v^n \rightharpoonup v$ weakly in $H^1(\omega;\mathbb{R}^3)$, $\varphi^n \to \varphi$ strongly in $W^{2,p}(\omega;\mathbb{R}^3)$ for all $p < +\infty$, and $\varphi^n \overset{*}{\rightharpoonup} \varphi$ weakly-$*$ in $W^{2,\infty}(\omega;\mathbb{R}^3)$, then $\gamma_{\alpha\beta}^n(v^n) \rightharpoonup \gamma_{\alpha\beta}(v)$ weakly in $L^2(\omega)$ and $\Upsilon_{\alpha\beta}^n(v^n) \rightharpoonup \Upsilon_{\alpha\beta}(v)$ weakly in $H^{-1}(\omega)$.*

Proof. First, since $a_\alpha^n \to a_\alpha$ strongly in $C^0(\bar{\omega};\mathbb{R}^3)$, it follows clearly that

$$\gamma_{\alpha\beta}^n(v^n) = \frac{1}{2}(\partial_\alpha v^n \cdot a_\beta^n + \partial_\beta v^n \cdot a_\alpha^n) \rightharpoonup \gamma_{\alpha\beta}(v) \text{ in } L^2(\omega).$$

The case of the change of curvature tensor is more intricate. We know that $\partial_{\alpha\beta} v^n \rightharpoonup \partial_{\alpha\beta} v$ in $H^{-1}(\omega;\mathbb{R}^3)$. Let θ be a test-function in $H_0^1(\omega)$. By the Sobolev embedding theorem, $\theta \in L^4(\omega)$, and, as $\partial_\alpha a_3^n \to \partial_\alpha a_3$ strongly in $L^4(\omega;\mathbb{R}^3)$ by (4.14) with $p = 4$, it follows that

$$\begin{cases} \theta a_3^n \to \theta a_3, \\ \partial_\alpha(\theta a_3^n) = (\partial_\alpha \theta) a_3^n + \theta \partial_\alpha a_3^n \to \partial_\alpha(\theta a_3), \end{cases} \text{strongly in } L^2(\omega;\mathbb{R}^3)$$

so that

$$\theta a_3^n \to \theta a_3 \text{ strongly in } H_0^1(\omega;\mathbb{R}^3).$$

Therefore,

$$\langle \partial_{\alpha\beta} v^n \cdot a_3^n, \theta\rangle = \langle \partial_{\alpha\beta} v^n, \theta a_3^n\rangle \to \langle \partial_{\alpha\beta} v \cdot a_3, \theta\rangle;$$

hence

$$\partial_{\alpha\beta} v^n \cdot a_3^n \rightharpoonup \partial_{\alpha\beta} v \cdot a_3 \text{ weakly in } H^{-1}(\omega).$$

Let us now deal with the other part of $\Upsilon^n_{\alpha\beta}(v^n)$. We have that $\partial_\rho v^n \rightharpoonup \partial_\rho v$ weakly in $L^2(\omega;\mathbb{R}^3)$, $\Gamma^{n,\rho}_{\alpha\beta} \to \Gamma^\rho_{\alpha\beta}$ strongly in $L^p(\omega)$, and $a_3^n \to a_3$ strongly in $W^{1,p}(\omega;\mathbb{R}^3)$ for all $p < +\infty$. It follows easily from this and Hölder's inequality that

$$\Gamma^{n,\rho}_{\alpha\beta} \partial_\rho v^n \cdot a_3^n \rightharpoonup \Gamma^\rho_{\alpha\beta} \partial_\rho v \cdot a_3 \text{ weakly in } L^q(\omega)\ \forall\ 1 < q < 2.$$

Now, by the two-dimensional Sobolev embedding theorem, we have $L^q(\omega) \hookrightarrow H^{-1}(\omega)$ for all $1 < q < 2$, hence the result. □

Let us now establish some uniform norm equivalence results.

Lemma 4.12. *There exist two constants* $0 < c < C < +\infty$ *independent of* n *such that for all* $v^n \in V_1^n$,

$$c\|v^n\|_{W^n} \le \left\{ \sum_{\alpha\beta} \left(\|\gamma^n_{\alpha\beta}(v^n)\|^2_{L^2(\omega;\mathbb{R}^3)} + \|\Upsilon^n_{\alpha\beta}(v^n)\|^2_{L^2(\omega;\mathbb{R}^3)} \right) \right\}^{1/2} \le C\|v^n\|_{W^n}. \tag{4.19}$$

Proof. The proof is essentially identical to that of Lemma 11 in [6] for a single chart. □

We also need uniform positive definiteness of the elasticity tensors. By assumption, for each midsurface, there exists a constant $\eta_n > 0$ such that for all symmetric tensors $\tau = (\tau_{\alpha\beta})$ and almost all $x \in \omega$, $a^{n,\alpha\beta\rho\sigma}(x)\tau_{\alpha\beta}\tau_{\rho\sigma} \ge \eta_n \tau_{\alpha\beta}\tau_{\alpha\beta}$. For instance, in the case of an isotropic material, this is a statement concerning the Lamé moduli μ and λ and not the geometry of the midsurface. We will concentrate here on the isotropic case.

Lemma 4.13. *There is constant* $\eta > 0$ *independent of* n *such that* $\eta_n \ge \eta$ *for all* n.

Proof. We know that $a^{n,\alpha\beta\rho\sigma}$ converge uniformly to $a^{\alpha\beta\rho\sigma}$. As η_n is the infimum of the quadratic form $a^{n,\alpha\beta\rho\sigma}(x)\tau_{\alpha\beta}\tau_{\rho\sigma}$ on the Cartesian product of the unit sphere of the space of symmetric tensors with $\bar{\omega}$, the result is clear. □

We now are in a position to establish uniform bounds for the various quantities of interest.

Lemma 4.14. *There is constant M independent of n such that*

$$\|u^n\|_{H^1(\omega;\mathbb{R}^3)} \le M, \qquad \|\partial_{\alpha\beta} u^n \cdot a_3^n\|_{L^2(\omega)} \le M. \tag{4.20}$$

Proof. Let us take $v^n = u^n$ as a test-function in the variational formulation of Koiter's problem (4.17). In view of Lemmas 4.12 and 4.13, we obtain

$$c\eta\sqrt{\delta}\min(e, e^3/12)\|u^n\|_{W^n}^2 \le \|\sqrt{a^n} f^n\|_{L^2(\omega;\mathbb{R}^3)}\|u^n\|_{L^2(\omega;\mathbb{R}^3)},$$

where δ is a uniform lower bound for $a^n(x)$. The lemma easily follows from the above estimate. □

Lemma 4.15. *There exists a subsequence* (*still denoted by*) u^n *and* $u \in V_1$ *such that*

$$u^n \rightharpoonup u \textit{ weakly in } H^1(\omega;\mathbb{R}^3) \quad \textit{and} \quad \partial_{\alpha\beta} u^n \cdot a_3^n \rightharpoonup \partial_{\alpha\beta} u \cdot a_3 \textit{ weakly in } L^2(\omega). \tag{4.21}$$

Proof. Because of the previous bounds, we can find a subsequence u^n, a function $u \in H_0^1(\omega;\mathbb{R}^3)$, and functions $\kappa_{\alpha\beta} \in L^2(\omega)$ such that

$$u^n \rightharpoonup u \text{ weakly in } H^1(\omega;\mathbb{R}^3) \quad \text{and} \quad \partial_{\alpha\beta} u^n \cdot a_3^n \rightharpoonup \kappa_{\alpha\beta} \text{ weakly in } L^2(\omega).$$

As in the proof of Lemma 4.11, we see that $\kappa_{\alpha\beta} = \partial_{\alpha\beta} u \cdot a_3$ so that $u \in W$. Moreover, $\partial_\alpha u^n \cdot a_3^n \rightharpoonup \partial_\alpha u \cdot a_3$ in $H^1(\omega)$ so that $\partial_\alpha u \cdot a_3 \in H_0^1(\omega)$ and therefore $u \in V_1$. □

Our next task is to identify the weak limit u of the above subsequence of solutions u^n as being the solution to the Koiter problem corresponding to the limit midsurface and loads.

Lemma 4.16. *The limit u is the unique solution to the following: Find $u \in V_1$ such that*

$$\forall v \in V_1, \quad \int_\omega e a^{\alpha\beta\rho\sigma}\left(\gamma_{\alpha\beta}(u)\gamma_{\rho\sigma}(v) + \frac{e^2}{12}\Upsilon_{\alpha\beta}(u)\Upsilon_{\rho\sigma}(v)\right)\sqrt{a}\,dx = \int_\omega f\cdot v\sqrt{a}\,dx. \tag{4.22}$$

The whole sequence is convergent.

Proof. By Theorem 4.4, we know that the spaces V_1^n and V_1 all share a common dense subspace, namely here, $\mathcal{D}(\omega;\mathbb{R}^3)$. Therefore, any $\psi \in \mathcal{D}(\omega;\mathbb{R}^3)$ is a legitimate test-function for all n, as well as for the eventual limit problem (4.22). Now

$$a^{n,\alpha\beta\rho\sigma}\gamma_{\rho\sigma}^n(\psi)\sqrt{a^n} \to a^{\alpha\beta\rho\sigma}\gamma_{\rho\sigma}(\psi)\sqrt{a} \text{ strongly in } L^2(\omega)$$

and

$$a^{n,\alpha\beta\rho\sigma}\Upsilon_{\rho\sigma}^n(\psi)\sqrt{a^n} \to a^{\alpha\beta\rho\sigma}\Upsilon_{\rho\sigma}(\psi)\sqrt{a} \text{ strongly in } L^2(\omega)$$

by Lemma 4.9. On the other hand,

$$\gamma_{\alpha\beta}^n(u^n) \rightharpoonup \gamma_{\alpha\beta}(u) \text{ weakly in } L^2(\omega)$$

and

$$\Upsilon_{\alpha\beta}^n(u^n) \rightharpoonup \Upsilon_{\alpha\beta}(u) \text{ weakly in } L^2(\omega)$$

by Lemma 4.11 and since Lemmas 4.12 and 4.14 imply that $\Upsilon_{\alpha\beta}^n(u^n)$ is bounded in $L^2(\omega)$. Therefore, we can pass to the limit as $n \to +\infty$ in problem (4.17) and obtain problem (4.22) for all test-functions $\psi \in \mathcal{D}(\omega;\mathbb{R}^3)$. The identification of u then follows from the fact that $\mathcal{D}(\omega;\mathbb{R}^3)$ is dense in V_1, viz. Theorem 4.4.

The solution of problem (4.22) is unique; therefore, the standard uniqueness argument shows that the whole sequence u^n converges, and not just a subsequence thereof. □

The final step in the proof of Theorem 4.10 consists in showing that all weak convergences are actually strong. This would be a straightforward matter if the test-function spaces did not depend on n. We would just take $u - u^n$ as a test-function. There is a slight twist here since $u \notin V_1^n$ and $u^n \notin V_1$ so that $u - u^n$ is a legitimate test-function neither for problem (4.17) nor for problem (4.22). We recast the problem in abstract form to circumvent this difficulty.

Let us be given a family of Hilbert spaces $(H^n, \|\cdot\|_n)_{n\in\mathbb{N}}$ and a Hilbert space $(H, \|\cdot\|)$ with the following properties:

(i) There is a subspace D common to all H^n and H such that D is dense in H.

(ii) For all $y \in D$, $\|y\|_n \to \|y\|$ as $n \to +\infty$.

Let us also be given a corresponding family of continuous (on their respective spaces) symmetric bilinear forms a^n and a and a family of continuous linear forms l^n and l with the following properties:

(iii) There exists a constant η independent of n such that

$$\forall y^n \in H^n, \quad a^n(y^n, y^n) \geq \eta \|y^n\|_n^2.$$

(iv) For all y, z in D, $a^n(y,z) \to a(y,z)$ and $l^n(y) \to l(y)$ when $n \to +\infty$.

Lemma 4.17. *Assume that hypotheses* (i)–(iv) *are satisfied, and let $x^n \in H^n$ and $x \in H$ be the solutions of the variational problems*

$$\forall y^n \in H^n, \quad a^n(x^n, y^n) = l^n(y^n) \quad \textit{and} \quad \forall y \in H, \quad a(x,y) = l(y).$$

If, in addition, $l^n(x^n) \to l(x)$, then

$$\|x^n\|_n \to \|x\| \quad \textit{when} \quad n \to +\infty. \tag{4.23}$$

Proof. For all $y \in D$, we have

$$a^n(x^n - y, x^n - y) = l^n(x^n - 2y) + a^n(y,y).$$

By assumption (iii), it follows that

$$\eta \|x^n - y\|_n^2 \leq l^n(x^n) - 2l^n(y) + a^n(y,y).$$

Letting n tend to $+\infty$, we obtain

$$\limsup_{n\to+\infty} \|x^n - y\|_n^2 \leq \frac{1}{\eta}(l(x) - 2l(y) + a(y,y))$$

by assumption (iv). Therefore,

$$\limsup_{n\to+\infty} \big| \|x^n\|_n - \|y\|_n \big| \leq \sqrt{\frac{1}{\eta}(l(x) - 2l(y) + a(y,y))}.$$

Now

$$\big| \|x^n\|_n - \|x\| \big| \leq \big| \|x^n\|_n - \|y\|_n \big| + \big| \|y\|_n - \|y\| \big| + \big| \|y\| - \|x\| \big|$$

so that, letting n tend to $+\infty$, we obtain

$$\limsup_{n\to+\infty}\big|\|x^n\|_n-\|x\|\big|\leq\sqrt{\frac{1}{\eta}(l(x)-2l(y)+a(y,y))}+\big|\|y\|-\|x\|\big|$$

for all $y\in D$, by assumption (ii). The lemma then results from assumption (i) and the fact that a and l are continuous on H. ∎

We can now apply Lemma 4.17 to our shell problem to complete the proof of Theorem 4.10.

Lemma 4.18. *We have*

(4.24) $u^n\to u$ *strongly in* $H^1(\omega;\mathbb{R}^3)$ *and* $\Upsilon^n_{\alpha\beta}(u^n)\to\Upsilon_{\alpha\beta}(u)$ *strongly in* $L^2(\omega)$.

Proof. By the weak convergence result of Lemma 4.15, we know that

$$\liminf_{n\to+\infty}\|u^n\|_{H^1(\omega;\mathbb{R}^3)}\geq\|u\|_{H^1(\omega;\mathbb{R}^3)}\quad\text{and}\quad\liminf_{n\to+\infty}\|\partial_{\alpha\beta}u^n\cdot a_3^n\|_{L^2(\omega)}\geq\|\partial_{\alpha\beta}u\cdot a_3\|_{L^2(\omega)}.$$

The Hilbert spaces V_1^n and V_1 and the bilinear and linear forms associated with the Koiter problems clearly satisfy the hypotheses of Lemma 4.17 with $D=\mathcal{D}(\omega;\mathbb{R}^3)$. Therefore,

$$\|u^n\|^2_{H^1(\omega;\mathbb{R}^3)}+\sum_{\alpha\beta}\|\partial_{\alpha\beta}u^n\cdot a_3^n\|^2_{L^2(\omega)}=\|u^n\|^2_{V_1^n}$$
$$\to\|u\|^2_{V_1}=\|u\|^2_{H^1(\omega;\mathbb{R}^3)}+\sum_{\alpha\beta}\|\partial_{\alpha\beta}u\cdot a_3\|^2_{L^2(\omega)},$$

which, together with the previous estimates, implies that

$$\|u^n\|_{H^1(\omega;\mathbb{R}^3)}\to\|u\|_{H^1(\omega;\mathbb{R}^3)}\quad\text{and}\quad\|\partial_{\alpha\beta}u^n\cdot a_3^n\|_{L^2(\omega)}\to\|\partial_{\alpha\beta}u\cdot a_3\|_{L^2(\omega)}.$$

The first convergence implies that $u^n\to u$ strongly in $H^1(\omega;\mathbb{R}^3)$, and the second convergence implies that $\partial_{\alpha\beta}u^n\cdot a_3^n\to\partial_{\alpha\beta}u\cdot a_3$ strongly in $L^2(\omega)$. Both facts imply that $\partial_\alpha u^n\cdot a_3^n\to\partial_\alpha u\cdot a_3$ strongly in $H^1(\omega)$; therefore, by the Sobolev embedding theorem and Lemma 4.9, $\Gamma^{n,\rho}_{\alpha,\beta}\partial_\rho u^n\cdot a_3^n\to\Gamma^{\rho}_{\alpha,\beta}\partial_\rho u\cdot a_3$ strongly in $L^2(\omega)$, which completes the proof. ∎

Remark 4.19. Note that the convergences established in Theorem 4.10 are quite natural in the sense that they imply the convergence of the displacements and of the associated strain and change of curvature tensors in their respective natural spaces. This in turn implies the strong L^2 convergence of the various stress resultants. ∎

Let us close the article with the final comparison between the classical formulation of Koiter's model and our formulation. Let us be given a sequence of charts φ^n as in Theorem 4.10. For any displacement $v\in V_1^n$, we denote by $v_i^n=v^n\cdot a_i^n$ the covariant components of v so that

$$v^n(x)=v_i^n(x)a^{n,i}(x).$$

Note that in some sense, in considering the covariant components as the basic unknown as is classically done, one mixes regularity issues concerning the displacement with regularity issues concerning the chart. This leads to the restrictive C^3 assumption that is made in the classical formulation. This remark may be illustrated by the following result.

THEOREM 4.20. *Let φ^n be a sequence of C^3-charts such that $\varphi^n \to \varphi$ in $W^{2,p}(\omega;\mathbb{R}^3)$ strong for all $1 < p < +\infty$ and $\varphi^n \overset{*}{\rightharpoonup} \varphi$ in $W^{2,\infty}(\omega;\mathbb{R}^3)$ weak-$*$ with φ piecewise C^3 and $\varphi \notin C^3(\omega;\mathbb{R}^3)$. Let f^n be a sequence of force resultant densities such that $f^n \to f$ in $L^2(\omega;\mathbb{R}^3)$. Let $(u_1^n, u_2^n, u_3^n) \in H_0^1(\omega) \times H_0^1(\omega) \times H_0^2(\omega)$ be the solution of the classical formulation of Koiter's problem as in* [4]. *Then, for all $n \in \mathbb{N}$,*

$$u^n(x) = u_i^n(x) a^{n,i}(x), \tag{4.25}$$

u^n tends to u in the sense of Theorem 4.10, *but u_3^n is generically unbounded in $H^2(\omega)$.*

Proof. The fact that (4.25) holds was already noted in [6]. Clearly, $u_3^n \to u_3$ in $L^2(\omega)$ by Theorem 4.10. If u_3^n was bounded in $H^2(\omega)$, this would thus imply that $u_3 \in H^2(\omega)$. This is not the case. As was already noted in [6], for a piecewise C^3-midsurface, the derivatives of the second fundamental form contain Dirac masses concentrated on the interfaces between the smooth parts of the shell. The condition $\partial_{\alpha\beta} u \cdot a_3 \in L^2(\omega)$ is equivalent to $\partial_\alpha u \cdot a_3 \in H^1(\omega)$, which means that the jump of $\partial_\alpha u \cdot a_3$ vanishes on each interface. In covariant components, this reads $[\partial_\alpha u_3 + b_\alpha^\rho u_\rho] = 0$, or, equivalently, $[\partial_\alpha u_3] = -[b_\alpha^\rho] u_\rho$, on each interface (with u_3 piecewise H^2), where b_α^ρ denote the mixed components of the second fundamental form. Since the jump of b_α^ρ is nonzero for some components, this will generically induce a jump on $\partial_\alpha u_3$. The normal component u_3 thus cannot be in $H^2(\omega)$. $\square$

Remark 4.21. The previous result indicates that a continuous dependence analysis similar to the present one would be difficult to carry out in the classical formulation. Some extra conditions would need to be imposed on the sequence of midsurfaces in order to obtain a uniform H^2-bound on u_3^n. $\square$

Example 4.22. Let us consider the example of a $W^{2,\infty}$-shell made of a plane part and a circular cylindrical part. We take $\omega = \{-\pi/2, \pi/2\} \times \{0, 1\}$ and

$$\varphi(x) = \begin{cases} (x_1, x_2, 0)^T & \text{for } x_1 \leq 0, \\ (\sin x_1, x_2, 1 - \cos x_1)^T & \text{for } x_1 > 0. \end{cases}$$

The midsurface of this shell is of class $W^{2,\infty}$ and has a curvature discontinuity across $x_1 = 0$. In particular, it is not C^3 and the classical formulation of Koiter's model is not applicable. It is easy to construct an explicit sequence of C^3-midsurfaces φ^n that converge in all $W^{2,p}$ and in $W^{2,\infty}$ weak-$*$ by using Hermite interpolation polynomials in the strip $\{0 \leq x_1 \leq 1/n\}$. The sequence of solutions associated with the sequence of interpolated midsurfaces falls into the classical framework. The limit midsurface is not of class C^3; thus the limit displacement requires our new formulation. The interface between the smooth parts is the line $\{x_1 = 0\}$. On this interface, all mixed components of the second fundamental form are continuous except b_1^1, and it is easy to see that $[b_1^1] = 1$. Therefore, we obtain that $[\partial_1 u_3] = -u_1$ (and $[\partial_2 u_3] = 0$). Since, in general, u_1 will not vanish on $\{x_1 = 0\}$, we see that u_3 is not in $H^2(\omega)$. This is not surprising if we remember that u_3 is a covariant component of u, hence a scalar product of u with the covariant basis vector a_3. The (lack of) regularity of a_3 is thus necessarily reflected in the degree of regularity of u_3. $\square$

Acknowledgments. The authors would like to thank V. Girault and L. R. Scott for fruitful discussions concerning Friedrichs's lemma and its extension up to the boundary.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Pure Appl. Math. 65, Academic Press, New York, London, 1975.

[2] C. Bardos, *Problèmes aux limites pour les équations aux dérivées partielles du premier ordre à coefficients réels; théorèmes d'approximation; application à l'équation de transport*, Ann. Sci. École Norm. Sup. (4), 3 (1970), pp. 185–233.

[3] M. Bernadou, *Méthodes d'éléments finis pour les problèmes de coques minces*, Rech. Math. Appl. 33, Masson, Paris, 1994.

[4] M. Bernadou and P. G. Ciarlet, *Sur l'ellipticité du modèle linéaire de coques de W.T. Koiter*, in Computing Methods in Sciences and Engineering, Lecture Notes in Econom. and Math. Systems 134, R. Glowinski and J.-L. Lions, eds., Springer-Verlag, Berlin, 1976, pp. 89–136.

[5] A. Blouza and H. Le Dret, *Existence et unicité pour le modèle de Koiter pour une coque peu régulière*, C. R. Acad. Sci. Paris Sér. I Math., 319 (1994), pp. 1127–1132.

[6] A. Blouza and H. Le Dret, *Existence and uniqueness for the linear Koiter model for shells with little regularity*, Quart. Appl. Math., 57 (1999), pp. 317–337.

[7] D. Chenais, *Un résultat de compacité d'un ensemble de points de* $\mathbb{R}^n$, C. R. Acad. Sci. Paris Sér. I Math., 277 (1973), pp. 905–907.

[8] P. G. Ciarlet, *Mathematical Elasticity. Vol.* III: *Theory of Shells*, North-Holland, Amsterdam, 2000.

[9] M. C. Delfour, *Intrinsic differential geometric methods in the asymptotic analysis of linear thin shells*, in Boundaries, Interfaces, and Transitions (Banff, AB, 1995), CRM Proc. Lecture Notes 13, AMS, Providence, RI, 1998, pp. 19–90.

[10] B. Desjardins, *A few remarks on ordinary differential equations*, Comm. Partial Differential Equations, 21 (1996), pp. 1667–1703.

[11] P. Destuynder and M. Salaün, *A mixed finite element for shell model with free edge boundary conditions. Part* 1. *The mixed variational formulation*, Comput. Methods Appl. Mech. Engrg., 120 (1995), pp. 195–217.

[12] R. J. DiPerna and P.-L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[13] D. E. Edmunds and W. D. Evans, *Spectral Theory and Differential Operators*, Oxford Math. Monogr., Oxford University Press, New York, 1987.

[14] L. E. Fraenkel, *On regularity of the boundary in the theory of Sobolev spaces*, Proc. London Math. Soc. (3), 39 (1979), pp. 385–427.

[15] V. Girault and L. R. Scott, *Analysis of a two-dimensional grade-two fluid model with a tangential boundary condition*, J. Math. Pures Appl., 78 (1999), pp. 981–1011.

[16] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monographs and Studies in Mathematics 24, Pitman, Boston, London, 1985.

[17] L. Hörmander, *Weak and strong extensions of differential operators*, Comm. Pure Appl. Math., 14 (1961), pp. 371–379.

[18] L. Hörmander, *The Analysis of Linear Partial Differential Operators.* III. *Pseudodifferential Operators*, Grundlehren Math. Wiss. 274, Springer-Verlag, Berlin, New York, 1985.

[19] N. Kerdid and P. Mato-Eiroa, *Approximation par éléments finis conformes d'un modèle de coques peu régulières*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 1335–1340.

[20] N. Kerdid and P. Mato-Eiroa, *Conforming finite element approximation for shells with little regularity*, Comput. Methods Appl. Mech. Engrg., 188 (2000), pp. 95–107.

[21] W. T. Koiter, *On the foundations of the linear theory of thin elastic shells.* I, II, Nederl. Akad. Wetensch. Proc. Ser. B, 73 (1970), pp. 169–195.

[22] J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.

[23] J.-P. Puel and M.-C. Roptin, *Lemme de Friedrichs, théorèmes de densité résultant du lemme de Friedrichs*, Graduate report under the supervision of C. Goulaouic, Diplôme d'Études Approfondies, Université de Rennes, Rennes, France, 1968.

[24] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton Math. Ser. 30, Princeton University Press, Princeton, NJ, 1970.

SIAM J. MATH. ANAL.
Vol. 33, No. 4, pp. 896–912

REDUCTION AND A CONCENTRATION-COMPACTNESS PRINCIPLE FOR ENERGY-CASIMIR FUNCTIONALS[*]

GERHARD REIN[†]

Abstract. Energy-Casimir functionals are a useful tool for the construction of steady states and the analysis of their nonlinear stability properties for a variety of conservative systems in mathematical physics. Recently, Y. Guo and the author employed them to construct stable steady states for the Vlasov–Poisson system in stellar dynamics, where the energy-Casimir functionals act on number density functions on phase space. In the present paper we construct natural, reduced functionals which act on mass densities on space and study compactness properties and the existence of minimizers in this context. This puts the techniques developed by Y. Guo and the author into a more general framework. We recover the concentration-compactness principle due to P.-L. Lions [*Ann. Inst. H. Poincaré Anal. Non Linéaire*, 1 (1984), pp. 109–145] in a more specific setting and connect our stability analysis with that of G. Wolansky [*Ann. Inst. H. Poincaré Anal. Non Linéaire*, 16 (1999), pp. 15–48].

Key words. energy-Casimir functionals, reduction, concentration-compactness principle, nonlinear stability, Vlasov–Poisson system

AMS subject classifications. 35A15, 35B35, 85A05

PII. P0036141001389275

1. Introduction.

The purpose of the present paper is to investigate the compactness properties and existence of minimizers of certain functionals which appear naturally in the stability analysis of various systems in kinetic theory. Given a large ensemble of particles which interact by gravitational attraction we consider energy-Casimir functionals which are defined on the space of phase space density functions, and certain reduced versions of these which are defined on the space of spatial density functions. This reduction procedure should put the techniques developed in [1, 2, 3, 4, 5, 11, 12] into a more general framework and make them applicable to problems outside kinetic theory. However, to be specific we start by recalling the Vlasov–Poisson system which describes the time evolution of a large ensemble of particles interacting by the gravitational field which they create collectively:

$$\partial_t f + v \cdot \partial_x f - \partial_x U \cdot \partial_v f = 0,$$

$$\triangle U = 4\pi\,\rho, \qquad \lim_{|x|\to\infty} U(t,x) = 0,$$

$$\rho(t,x) = \int f(t,x,v)dv.$$

Here the dynamic variable is the number density $f = f(t,x,v)$ of the ensemble in phase space, $x, v \in \mathbb{R}^3$ denote position and velocity, $\rho = \rho(t,x)$ is the spatial mass

[*] Received by the editors May 14, 2001; accepted for publication (in revised form) August 10, 2001; published electronically December 28, 2001. This research was supported by the Erwin Schrödinger International Institute for Mathematical Physics in Vienna and by the Wittgenstein Award 2000 of Peter A. Markowich, Vienna.

http://www.siam.org/journals/sima/33-4/38927.html

[†] Institut für Mathematik der Universität Wien, Strudlhofgasse 5, 1090 Vienna, Austria (gerhard.rein@univie.ac.at).

density induced by f, and $U = U(t,x)$ is the induced gravitational potential. It is straightforward to check that

$$\iint Q(f(t,x,v))\,dv\,dx + \frac{1}{2}\iint |v|^2 f(t,x,v)\,dv\,dx - \frac{1}{2}\iint \frac{\rho(t,x)\,\rho(t,y)}{|x-y|}\,dx\,dy$$

is conserved along solutions for any suitable scalar function Q. The first part, which is conserved by itself, is a so-called Casimir functional, the second is the kinetic, and the third part is the potential energy of the system. When viewed as a functional on phase space densities $f = f(x,v) \geq 0$ we denote this functional by $\mathcal{H}_{\mathcal{C}}$. It is fairly straightforward to see that any minimizer of this functional subject to the constraint

$$\iint f(x,v)\,dv\,dx = M$$

with prescribed total mass $M > 0$ is a steady state of the Vlasov–Poisson system. It is far less obvious that such minimizers exist and that they are nonlinearly stable. When analyzing the minimization problem

$$\mathcal{H}_{\mathcal{C}}(f_0) = \inf\left\{\mathcal{H}_{\mathcal{C}}(f) | f \geq 0, \iint f\,dv\,dx = M\right\}, \tag{1.1}$$

one needs to make sure that one can pass to the limit in the (quadratic) potential energy along a minimizing sequence. Obviously, the potential energy is not a functional of f itself but of the induced spatial density ρ, and the crucial question is how, along a minimizing sequence, the spatial density can or cannot split into parts or spread uniformly in space. This was analyzed in the context of the Vlasov–Poisson system and with various variations in [1, 2, 3, 4, 5, 11, 12].

In the present paper we want to bring out the basic mechanism more clearly and in a framework not restricted to kinetic theory. To this end we construct in the next section a reduced version $\mathcal{H}_{\mathcal{C}}^r$ of the energy-Casimir functional $\mathcal{H}_{\mathcal{C}}$, which will be defined on spatial densities ρ:

$$\mathcal{H}_{\mathcal{C}}^r(\rho) = \int \Phi(\rho(x))\,dx - \frac{1}{2}\iint \frac{\rho(x)\,\rho(y)}{|x-y|}\,dx\,dy$$

with Φ a function determined by Q, which is convex if Q is convex. Then we explore the relation between the variational problems

$$\mathcal{H}_{\mathcal{C}}^r(\rho_0) = \inf\left\{\mathcal{H}_{\mathcal{C}}^r(\rho) | \rho \geq 0, \int \rho\,dx = M\right\} \tag{1.2}$$

and (1.1); in particular we will show how a minimizer of the reduced problem (1.2) induces a minimizer of (1.1). In the third section we reformulate the techniques developed for (1.1) in the framework of (1.2) and obtain the existence of a minimizer ρ_0 under appropriate conditions on Φ. In particular, we prove the essential part of the concentration-compactness principle due to P.-L. Lions [9] by a more direct method based on scaling and splitting. In the last section we discuss the role of symmetries in the problem and point out some applications and extensions of our results. An example of a function Φ which satisfies all the necessary assumptions is $\Phi(\rho) = \rho^{1+1/n}$ with $0 < n < 3$. In this case the potential U_0 induced by a minimizer ρ_0 is a solution of the semilinear elliptic problem

$$\triangle U_0 = (E_0 - U_0)_+^n, \qquad \lim_{|x|\to\infty} U_0(x) = 0,$$

where $(\cdot)_+$ denotes the positive part and E_0 is some constant. This equation is sometimes referred to as the Emden–Fowler equation and appears naturally in the study of self-gravitating fluid balls. Throughout this paper we restrict ourselves to the case of space dimension 3; extending these techniques to other space dimensions by adjusting various exponents is easy. We remark that the use of energy-Casimir functionals for questions of stability was discussed in a very broad context in [7].

2. Reduction of energy-Casimir functionals. For a measurable function $f = f(x,v)$ we define

$$\rho_f(x) := \int f(x,v)\,dv, \qquad x \in \mathbb{R}^3,$$

and

$$U_f := -\rho_f * \frac{1}{|\cdot|}.$$

Next we define

$$E_{\rm kin}(f) := \frac{1}{2}\iint |v|^2 f(x,v)\,dv\,dx,$$
$$E_{\rm pot}(f) := -\frac{1}{8\pi}\int |\nabla U_f(x)|^2 dx = -\frac{1}{2}\iint \frac{\rho_f(x)\rho_f(y)}{|x-y|}dx\,dy$$

($E_{\rm pot}$ can equally well be viewed as a functional of ρ instead of f), and

$$\mathcal{H}_{\mathcal{C}}(f) := \mathcal{C}(f) + E_{\rm kin}(f) + E_{\rm pot}(f),$$

where

$$\mathcal{C}(f) := \iint Q(f(x,v))\,dv\,dx,$$

and Q is a given function satisfying the following assumption.

Assumption on Q. $Q \in C^1([0,\infty[)$ is strictly convex, $Q(0) = Q'(0) = 0$, and $Q(f)/f \to \infty,\ f \to \infty$.

In particular, this implies that $Q \geq 0$ and $Q' : [0,\infty[\to [0,\infty[$ is one-to-one and onto.

We study the following variational problem: Minimize $\mathcal{H}_{\mathcal{C}}$ over the set

$$(2.1)\ \mathcal{F}_M := \left\{ f \in L^1_+(\mathbb{R}^6) \mid \mathcal{C}(f) + E_{\rm kin}(f) < \infty,\ \rho_f \in L^{6/5}(\mathbb{R}^3),\ \iint f = M \right\},$$

where $M > 0$ is prescribed and $L^1_+(\mathbb{R}^6)$ denotes the set of a.e. nonnegative functions in $L^1(\mathbb{R}^6)$. Note that since $\rho_f \in L^{6/5}(\mathbb{R}^3)$ the convolution defining U_f exists in $L^6(\mathbb{R}^3)$ with $\nabla U_f \in L^2(\mathbb{R}^3)$ according to the extended Young's inequality, and the potential energy of ρ_f is finite.

In order to guarantee the existence of a minimizer we will require additional growth conditions on Q to be introduced later; at the moment $E_{\rm pot}(f)$ could be minus infinity for $f \in \mathcal{F}_M$. A typical example of a function for which there exists a minimizer is $Q(f) = f^{1+1/k}$ with $0 < k < 3/2$.

To obtain a reformulation in terms of spatial densities ρ which captures the essential properties of this variational problem we proceed as follows. For $r \geq 0$ we define

$$\mathcal{G}_r := \left\{ g \in L^1_+(\mathbb{R}^3) | \int \left(\frac{1}{2}|v|^2 g(v) + Q(g(v)) \right) dv < \infty, \ \int g(v)\, dv = r \right\} \tag{2.2}$$

and

$$\Phi(r) := \inf_{g \in \mathcal{G}_r} \int \left(\frac{1}{2}|v|^2 g(v) + Q(g(v)) \right) dv. \tag{2.3}$$

In addition to the variational problem of minimizing $\mathcal{H}_\mathcal{C}$ over the set $\mathcal{F}_M$ we consider the problem of minimizing the functional

$$\mathcal{H}^r_\mathcal{C}(\rho) := \int \Phi(\rho(x))\, dx + E_{\rm pot}(\rho) \tag{2.4}$$

over the set

$$\mathcal{F}^r_M := \left\{ \rho \in L^{6/5} \cap L^1_+(\mathbb{R}^3) \mid \int \Phi(\rho(x))\, dx < \infty, \ \int \rho(x)\, dx = M \right\}. \tag{2.5}$$

The relation between the minimizers of $\mathcal{H}_\mathcal{C}$ and $\mathcal{H}^r_\mathcal{C}$ is the main theme of this section, and a remark on how we passed from $\mathcal{H}_\mathcal{C}$ to $\mathcal{H}^r_\mathcal{C}$ can be found at the end of the section.

Theorem 2.1.

(a) *For every function $f \in \mathcal{F}_M$,*

$$\mathcal{H}_\mathcal{C}(f) \geq \mathcal{H}^r_\mathcal{C}(\rho_f),$$

and if $f = f_0$ is a minimizer of $\mathcal{H}_\mathcal{C}$ over $\mathcal{F}_M$, then equality holds.

(b) *Let $\rho_0 \in \mathcal{F}^r_M$ be a minimizer of $\mathcal{H}^r_\mathcal{C}$ with induced potential U_0. Then there exists a Lagrange multiplier $E_0 \in \mathbb{R}$ such that a.e.*

$$\rho_0 = \begin{cases} (\Phi')^{-1}(E_0 - U_0), & U_0 < E_0, \\ 0, & U_0 \geq E_0. \end{cases}$$

Denote by

$$E = E(x,v) := \frac{1}{2}|v|^2 + U_0(x)$$

the energy of a particle at position x with velocity v, and define

$$f_0 := \begin{cases} (Q')^{-1}(E_0 - E), & E < E_0, \\ 0, & E \geq E_0. \end{cases}$$

Then $f_0 \in \mathcal{F}_M$ is a minimizer of $\mathcal{H}_\mathcal{C}$.

(c) *Now* assume *that $\mathcal{H}^r_\mathcal{C}$ has at least one minimizer in $\mathcal{F}^r_M$. Then the following holds: If $f_0 \in \mathcal{F}_M$ is a minimizer of $\mathcal{H}_\mathcal{C}$, then $\rho_0 := \rho_{f_0} \in \mathcal{F}^r_M$ is a minimizer of $\mathcal{H}^r_\mathcal{C}$. This map is one-to-one and onto between the sets of minimizers of $\mathcal{H}_\mathcal{C}$ in $\mathcal{F}_M$ and of $\mathcal{H}^r_\mathcal{C}$ in $\mathcal{F}^r_M$, respectively, and is the inverse of the map $\rho_0 \mapsto f_0$ described in* (b).

Remark. This theorem does not exclude the possibility that $\mathcal{H}_\mathcal{C}$ has a minimizer but $\mathcal{H}_\mathcal{C}^r$ has none. In the next section we show that under appropriate assumptions on Φ the reduced functional $\mathcal{H}_\mathcal{C}^r$ does have a minimizer, and then the theorem guarantees that we recover all minimizers of $\mathcal{H}_\mathcal{C}$ in $\mathcal{F}_M$ by "lifting" the ones of $\mathcal{H}_\mathcal{C}^r$ as described in (b).

Before we prove this theorem, we investigate the relation between Q and Φ; for a function $h : \mathbb{R} \to]-\infty, \infty]$ we denote by

$$h^*(\lambda) := \sup_{r \in \mathbb{R}} (\lambda\, r - h(r))$$

its Legendre transform. Some of the results of the lemma below will be relevant for the next section.

Lemma 2.2. *Let Q be as specified above, let Φ be defined by* (2.2), (2.3), *and extend both functions by $+\infty$ to the interval* $]-\infty, 0[$.

(a) *For $\lambda \in \mathbb{R}$,*

$$\Phi^*(\lambda) = \int Q^*\left(\lambda - \frac{1}{2}|v|^2\right) dv,$$

and, in particular, $Q^*(\lambda) = 0 = \Phi^*(\lambda)$ for $\lambda < 0$.

(b) *$\Phi \in C^1([0,\infty[)$ is strictly convex, and $\Phi(0) = \Phi'(0) = 0$.*

(c) *Let $k > 0$ and $n = k + 3/2$. As in the rest of the paper, constants denoted by C are positive, may depend on Q or M, and may change from line to line (or within one line).*

(i) *If $Q(f) = C\, f^{1+1/k},\ f \geq 0$, then $\Phi(\rho) = C\, \rho^{1+1/n},\ \rho \geq 0$.*

(ii) *If $Q(f) \geq C\, f^{1+1/k},\ f \geq 0$ large, then $\Phi(\rho) \geq C\, \rho^{1+1/n},\ \rho \geq 0$ large.*

(iii) *If $Q(f) \leq C\, f^{1+1/k},\ f \geq 0$ small, then $\Phi(\rho) \leq C\, \rho^{1+1/n},\ \rho \geq 0$ small.*

If the restriction to large, respectively, small values of f can be dropped, then the corresponding restriction for ρ can be dropped as well.

Proof. By definition,

$$\begin{aligned}
\Phi^*(\lambda) &= \sup_{r \geq 0} \left[\lambda\, r - \inf_{g \in \mathcal{G}_r} \int \left(\frac{1}{2}|v|^2 g(v) + Q(g(v))\right) dv\right] \\
&= \sup_{r \geq 0} \sup_{g \in \mathcal{G}_r} \int \left[\left(\lambda - \frac{1}{2}|v|^2\right) g(v) - Q(g(v))\right] dv \\
&= \sup_{g \in L^1_+(\mathbb{R}^3)} \int \left[\left(\lambda - \frac{1}{2}|v|^2\right) g(v) - Q(g(v))\right] dv \\
&= \int \sup_{y \geq 0} \left[\left(\lambda - \frac{1}{2}|v|^2\right) y - Q(y)\right] dv = \int Q^*\left(\lambda - \frac{1}{2}|v|^2\right) dv.
\end{aligned}$$

As to the last-but-one equality, observe that both sides are obviously zero for $\lambda \leq 0$. If $\lambda > 0$ then for any $g \in L^1_+(\mathbb{R}^3)$,

$$\int \left[\left(\lambda - \frac{1}{2}|v|^2\right) g(v) - Q(g(v))\right] dv \leq \int \sup_{y \geq 0} \left[\left(\lambda - \frac{1}{2}|v|^2\right) y - Q(y)\right] dv.$$

If $|v| \geq \sqrt{2\lambda}$ then $\sup_{y \geq 0} [\cdots] = 0$, and for $|v| < \sqrt{2\lambda}$ the supremum of the term in brackets is attained at $y = y_v := (Q')^{-1}\left(\lambda - \frac{1}{2}|v|^2\right)$. Thus with

$$g_0(v) := \begin{cases} y_v, & |v| < \sqrt{2\lambda}, \\ 0, & |v| \geq \sqrt{2\lambda}, \end{cases}$$

we have

$$\begin{aligned}\int \sup_{y\geq 0}\left[\left(\lambda-\frac{1}{2}|v|^2\right)y-Q(y)\right]dv &= \int\left[\left(\lambda-\frac{1}{2}|v|^2\right)g_0(v)-Q(g_0(v))\right]dv\\ &\leq \sup_{g\in L^1_+(\mathbb{R}^3)}\int\left[\left(\lambda-\frac{1}{2}|v|^2\right)g(v)-Q(g(v))\right]dv,\end{aligned}$$

and part (a) is established.

Since Q is strictly convex and lower semicontinuous as a function on $\mathbb{R}$ with $\lim_{|f|\to\infty} Q(f)/|f|\to\infty$, $Q^*\in C^1(\mathbb{R})$; cf. [10, Prop. 2.4]. Obviously, $Q^*(\lambda)=0$ for $\lambda\leq 0$; in particular, $(Q^*)'(0)=0$. Also, $(Q^*)'$ is strictly increasing on $[0,\infty[$ since Q' is strictly increasing on $[0,\infty[$ with range $[0,\infty[$. Since for $|\lambda|<\lambda_0$ with $\lambda_0>0$ fixed the integral in the formula for Φ^* extends over a compact set we may differentiate under the integral sign to conclude that $\Phi^*\in C^1(\mathbb{R})$ with derivative strictly increasing on $[0,\infty[$. This in turn implies the assertion of part (b).

Finally, $Q(f)\geq C\,f^{1+1/k},\ f\geq 0$ large, implies that $Q(f)\geq C\,f^{1+1/k}-C',\ f\geq 0$. Thus

$$Q^*(\lambda)\leq C'+\sup_{f\geq 0}\left(f\lambda-C\,f^{1+1/k}\right)=C'+\frac{1}{1+k}\left(\frac{k}{C_Q\,(1+k)}\right)^k\lambda^{1+k},\qquad \lambda\geq 0,$$

and

$$\begin{aligned}\Phi^*(\lambda)&\leq C\lambda^{3/2}+C\int_{|v|\leq\sqrt{2\lambda}}\left(\lambda-\frac{1}{2}|v|^2\right)^{1+k}dv=C\lambda^{3/2}+C\int_0^\lambda E^{1+k}\sqrt{\lambda-E}\,dE\\ &=C'+C\,\lambda^{k+5/2}=C'+C\lambda^{1+n},\qquad \lambda\geq 0.\end{aligned}$$

This in turn yields the assertion on Φ in (c)(ii). The assertion in (c)(i) is now obvious. As to (c)(iii) note first that for $\lambda\geq 0$ and small the corresponding supremum is attained at small f's, and thus

$$Q^*(\lambda)\leq\sup_{f\geq 0}\left(\lambda\,f-C\,f^{1+1/k}\right)=C\lambda^{1+k}.$$

Thus still for $\lambda\geq 0$ small, $\Phi^*(\lambda)\geq C\lambda^{1+n}$, which in turn implies the assertion for Φ. □

We now prove the theorem above.

Proof of Theorem 2.1. *Proof of the inequality in part* (a). For $\rho\in\mathcal{F}^r_M$ define

$$\mathcal{F}_\rho:=\{f\in\mathcal{F}_M|\rho_f=\rho\}. \tag{2.6}$$

Clearly, for $\rho=\rho_f$ with $f\in\mathcal{F}_M$,

$$\begin{aligned}\mathcal{C}(f)+E_{\mathrm{kin}}(f)&\geq\inf_{\tilde f\in\mathcal{F}_\rho}(\mathcal{C}(\tilde f)+E_{\mathrm{kin}}(\tilde f))\\ &\geq\inf_{\tilde f\in\mathcal{F}_\rho}\int\left[\inf_{g\in\mathcal{G}_{\rho(x)}}\int\left(\frac{1}{2}|v|^2g(v)+Q(g(v))\right)dv\right]dx\\ &=\int\left[\inf_{g\in\mathcal{G}_{\rho(x)}}\int\left(\frac{1}{2}|v|^2g(v)+Q(g(v))\right)dv\right]dx\\ &=\int\Phi(\rho(x))\,dx,\end{aligned} \tag{2.7}$$

and the inequality in part (a) is established.

An intermediate assertion. We claim that if $f \in \mathcal{F}_M$ is such that up to sets of measure zero,

$$\left\{ \begin{array}{rl} Q'(f) = E_0 - E > 0, & \text{where } f > 0, \\ E_0 - E \leq 0, & \text{where } f = 0, \end{array} \right. \tag{2.8}$$

with E defined as in (b) but with U_f instead of U_0 and E_0 a constant, then equality holds in part (a).

To prove this, observe that since Q is convex, we have for a.e. $x \in \mathbb{R}^3$ and every $g \in \mathcal{G}_{\rho_f(x)}$,

$$\begin{aligned} \frac{1}{2}|v|^2 g(v) + Q(g(v)) \geq{}& \frac{1}{2}|v|^2 f(x,v) + Q(f(x,v)) \\ &+ \left(\frac{1}{2}|v|^2 + Q'(f(x,v))\right)(g(v) - f(x,v)) \text{ a.e.} \end{aligned}$$

Now by (2.8),

$$\begin{aligned} \int \left(\frac{1}{2}|v|^2 + Q'(f)\right)(g - f)\,dv &= \int_{\{f>0\}} \dots + \int_{\{f=0\}} \dots \\ &= (E_0 - U_f(x)) \int_{\{f>0\}} (g - f)\,dv + \int_{\{f=0\}} \frac{1}{2}|v|^2 g\,dv \\ &= -(E_0 - U_f(x)) \int_{\{f=0\}} (g - f)\,dv + \int_{\{f=0\}} \frac{1}{2}|v|^2 g\,dv \\ &= \int_{\{f=0\}} (E - E_0)\, g\,dv \geq 0; \end{aligned}$$

observe that $g \geq 0$ and $\int g\,dv = \int f\,dv$ so $\int (g - f)\,dv = 0$. Thus we see that

$$\begin{aligned} \Phi(\rho_f(x)) &\geq \int \left(\frac{1}{2}|v|^2 f + Q(f)\right) dv \\ &\geq \inf_{g \in \mathcal{G}_{\rho_f(x)}} \int \left(\frac{1}{2}|v|^2 g + Q(g)\right) dv = \Phi(\rho_f(x)) \text{ a.e.,} \end{aligned}$$

and the proof of our intermediate assertion is complete.

Proof of the equality assertion in (a). If $f_0 \in \mathcal{F}_M$ is a minimizer of $\mathcal{H}_\mathcal{C}$ then the Euler–Lagrange equation of the minimization problem implies that (2.8) holds for some Lagrange multiplier E_0; this can be proved as in [5, Thm. 2]. Thus equality holds in (a) by the intermediate assertion, and the proof of part (a) is complete.

Proof of part (b). Let $\rho_0 \in \mathcal{F}_M^r$ be a minimizer of $\mathcal{H}_\mathcal{C}^r$. Then the Euler–Lagrange equation yields the relation between ρ_0 and U_0. Let f_0 be defined as in (b). Then up to sets of measure zero,

$$\begin{aligned} \int f_0(x,v)\,dv &= \int_{|v| \leq \sqrt{2(E_0 - U_0(x))}} (Q')^{-1}\left(E_0 - U_0(x) - \frac{1}{2}|v|^2\right) dv \\ &= (\Phi^*)'(E_0 - U_0(x)) = (\Phi')^{-1}(E_0 - U_0(x)) = \rho_0(x), \end{aligned}$$

where $U_0(x) < E_0$, and both sides are zero where $U_0(x) \geq E_0$. Thus $\rho_0 = \rho_{f_0}$, in particular, $f_0 \in \mathcal{F}_M$. By definition, f_0 satisfies the Euler–Lagrange relation (2.8) and thus by our intermediate assertion $\mathcal{H}_\mathcal{C}(f_0) = \mathcal{H}_\mathcal{C}^r(\rho_0)$. Therefore, again by part (a),

$$\mathcal{H}_\mathcal{C}(f) \geq \mathcal{H}_\mathcal{C}^r(\rho_f) \geq \mathcal{H}_\mathcal{C}^r(\rho_0) = \mathcal{H}_\mathcal{C}(f_0), \qquad f \in \mathcal{F}_M,$$

so that f_0 is a minimizer of $\mathcal{H}_\mathcal{C}$, and the proof of part (b) is complete.

Proof of part (c). Assume that $\mathcal{H}^r_\mathcal{C}$ has a minimizer $\rho_0 \in \mathcal{F}^r_M$ and define f_0 as above. Then part (a), the fact that each $\rho \in \mathcal{F}^r_M$ can be written as $\rho = \rho_f$ for some $f \in \mathcal{F}_M$, and our intermediate assertion imply that

$$\begin{aligned}\inf_{f\in\mathcal{F}_M} \mathcal{H}_\mathcal{C}(f) &\geq \inf_{f\in\mathcal{F}_M} \mathcal{H}^r_\mathcal{C}(\rho_f) = \inf_{\rho\in\mathcal{F}^r_M} \mathcal{H}^r_\mathcal{C}(\rho)\\ &= \mathcal{H}^r_\mathcal{C}(\rho_0) = \mathcal{H}_\mathcal{C}(f_0) \geq \inf_{f\in\mathcal{F}_M} \mathcal{H}_\mathcal{C}(f).\end{aligned} \tag{2.9}$$

Now take any minimizer $g_0 \in \mathcal{F}_M$ of $\mathcal{H}_\mathcal{C}$. Then by (2.9) and part (a),

$$\inf_{\rho\in\mathcal{F}^r_M} \mathcal{H}^r_\mathcal{C}(\rho) = \inf_{f\in\mathcal{F}_M} \mathcal{H}_\mathcal{C}(f) = \mathcal{H}_\mathcal{C}(g_0) = \mathcal{H}^r_\mathcal{C}(\rho_{g_0}),$$

that is, $\rho_{g_0} \in \mathcal{F}^r_M$ minimizes $\mathcal{H}^r_\mathcal{C}$, and the proof of part (c) is complete. □

Remark. If we define an intermediate functional

$$\mathcal{P}(\rho) := \inf_{f\in\mathcal{F}_\rho} \iint \left(\frac{1}{2}|v|^2 f(x,v) + Q(f(x,v))\right) dv\, dx$$

with $\mathcal{F}_\rho$ as defined in (2.6), then (2.7) shows that

$$\mathcal{H}_\mathcal{C}(f) \geq \mathcal{P}(\rho_f) + E_{\mathrm{pot}}(\rho_f) \geq \int \Phi(\rho_f(x))\, dx + E_{\mathrm{pot}}(\rho_f) = \mathcal{H}^r_\mathcal{C}(\rho_f)$$

with equality for minimizers. Note that $\mathcal{P}(\rho)$ is obtained by minimizing the positive contribution to $\mathcal{H}_\mathcal{C}$, which also happens to be the part depending on phase space densities f directly, over all f's which generate a given spatial density ρ. Then in a second step one minimizes for each point x over all functions $g = g(v)$ which have as integral the value $\rho(x)$.

These constructions are borrowed from [14] where they appear for the special case $Q(f) = f^{1+1/k}$. In [14] the resulting functional of ρ is investigated under the assumption of spherical symmetry by rewriting it as a functional of $m_\rho(r) := 4\pi \int_0^r s^2 \rho(s)\, ds$ where $r := |x|$. While minimizers of the present variational problems are spherically symmetric a posteriori, the a priori restriction to spherical symmetry implies that any stability result derived from their minimizing property is restricted to spherically symmetric perturbations, which is undesirable. Moreover, in the last section we will comment on some extensions of the present techniques to situations where the minimizers are not spherically symmetric.

3. Concentration-compactness principle and existence of minimizers.

In this section we prove a concentration-compactness principle that will yield a solution to the following variational problem: Minimize the functional

$$\mathcal{H}^r_\mathcal{C}(\rho) := \int \Phi(\rho(x))\, dx + E_{\mathrm{pot}}(\rho)$$

over the set

$$\mathcal{F}^r_M := \left\{\rho \in L^1_+(\mathbb{R}^3) \mid \int \Phi(\rho) < \infty,\ \int \rho = M\right\} \tag{3.1}$$

for $M > 0$ given and Φ satisfying the following.

Assumptions on Φ. $\Phi \in C^1([0,\infty[)$, $\Phi(0) = 0 = \Phi'(0)$, and

($\Phi 1$) Φ is strictly convex;
($\Phi 2$) $\Phi(\rho) \geq C\rho^{1+1/n}$, $\rho \geq 0$ large, with $0 < n < 3$;
($\Phi 3$) $\Phi(\rho) \leq C\rho^{1+1/n'}$, $\rho \geq 0$ small, with $0 < n' < 3$.

Note that Lemma 2.2 tells us that the function Φ, which we constructed from a given Q in section 2, has these properties, provided Q satisfies the growth conditions corresponding to ($\Phi 2$) and ($\Phi 3$). The aim of this section is to prove the following result.

Theorem 3.1. *The functional $\mathcal{H}^r_\mathcal{C}$ is bounded from below on $\mathcal{F}^r_M$. Let $(\rho_i) \subset \mathcal{F}^r_M$ be a minimizing sequence of $\mathcal{H}^r_\mathcal{C}$. Then there exists a sequence of shift vectors $(a_i) \subset \mathbb{R}^3$ and a subsequence, again denoted by (ρ_i), such that for any $\epsilon > 0$ there exists $R > 0$ with*

$$\int_{a_i+B_R} \rho_i(x)\,dx \geq M - \epsilon, \qquad i \in \mathbb{N},$$

$$T\rho_i := \rho_i(\cdot + a_i) \rightharpoonup \rho_0 \text{ weakly in } L^{1+1/n}(\mathbb{R}^3), \qquad i \to \infty,$$

and

$$\int_{B_R} \rho_0 \geq M - \epsilon.$$

Finally,

$$\nabla U_{T\rho_i} \to \nabla U_0 \text{ strongly in } L^2(\mathbb{R}^3), \qquad i \to \infty,$$

and $\rho_0 \in \mathcal{F}^r_M$ is a minimizer of $\mathcal{H}^r_\mathcal{C}$.

Here and in the following we denote for $0 < R < S \leq \infty$,

$$B_R := \{x \in \mathbb{R}^3 | |x| \leq R\},$$
$$B_{R,S} := \{x \in \mathbb{R}^3 | R \leq |x| < S\}.$$

We split our argument into a series of lemmas. The first thing to note is that $\mathcal{H}^r_\mathcal{C}$ is bounded from below on $\mathcal{F}^r_M$.

Lemma 3.2. *Under the above assumptions on Φ,*

$$\mathcal{H}^r_\mathcal{C}(\rho) \geq \int \Phi(\rho)\,dx - C - C\left(\int \Phi(\rho)\,dx\right)^{n/3}, \qquad \rho \in \mathcal{F}^r_M,$$

in particular,

$$h^r_M := \inf_{\mathcal{F}^r_M} \mathcal{H}^r_\mathcal{C} > -\infty.$$

Proof. By the extended Young's inequality, interpolation, and assumption ($\Phi 2$),

$$-E_{\text{pot}}(\rho) \leq C\|\rho\|^2_{6/5} \leq C\|\rho\|_1^{(5-n)/3}\|\rho\|_{1+1/n}^{(n+1)/3}$$
$$\leq C + C\left(\int \Phi(\rho)\,dx\right)^{n/3}, \qquad \rho \in \mathcal{F}^r_M.$$

Since $n < 3$, $\mathcal{H}^r_\mathcal{C}$ is bounded from below on $\mathcal{F}^r_M$. $\square$

Corollary 3.3. *Any minimizing sequence of $\mathcal{H}_\mathcal{C}^r$ in $\mathcal{F}_M^r$ is bounded in $L^{1+1/n}(\mathbb{R}^3)$ and therefore has a subsequence which converges weakly in $L^{1+1/n}(\mathbb{R}^3)$.*

Proof. By Lemma 3.2, $\int \Phi(\rho)$ is bounded along any minimizing sequence. The assertion follows by (Φ2) and the fact that $\int \rho = M$ for $\rho \in \mathcal{F}_M^r$. □

Note that the estimates above show that the definition (3.1) coincides with our earlier definition for the set $\mathcal{F}_M^r$. We also see that the assumption (Φ2) is quite natural. Next we prove a splitting estimate which will show that along a minimizing sequence the mass cannot vanish.

Lemma 3.4. *Let $\rho \in \mathcal{F}_M^r$. Then*

$$\sup_{a\in\mathbb{R}^3} \int_{a+B_R} \rho(x)\,dx \geq \frac{1}{RM}\left(-2E_{\text{pot}}(\rho) - \frac{M^2}{R} - \frac{C\|\rho\|^2_{1+1/n}}{R^{(5-n)/(n+1)}}\right), \qquad R > 1.$$

Proof. We split the potential energy as follows:

$$\begin{aligned} -2E_{\text{pot}}(\rho) &= \iint_{|x-y|\leq 1/R} \frac{\rho(x)\,\rho(y)}{|x-y|}\,dx\,dy + \iint_{1/R<|x-y|<R} \cdots + \iint_{R\geq|x-y|} \cdots \\ &=: I_1 + I_2 + I_3. \end{aligned}$$

By Hölder's inequality and Young's inequality,

$$\begin{aligned} I_1 &\leq \|\rho\|_{1+1/n} \|\rho * (\mathbf{1}_{B_{1/R}} 1/|\cdot|)\|_{n+1} \leq \|\rho\|^2_{1+1/n} \|\mathbf{1}_{B_{1/R}} 1/|\cdot|\|_{(n+1)/2} \\ &\leq C\,\|\rho\|^2_{1+1/n} R^{-(5-n)/(n+1)}; \end{aligned}$$

here $\mathbf{1}_S$ denotes the indicator function of the set $S \subset \mathbb{R}^3$. The estimates for I_2 and I_3 are straightforward:

$$I_2 \leq R \iint_{|x-y|\leq R} \rho(x)\,\rho(y)\,dx\,dy \leq M\,R \sup_{a\in\mathbb{R}^3} \int_{a+B_R} \rho(x)\,dx,$$

and

$$I_3 \leq R^{-1} M^2.$$

Putting these estimates together yields the assertion. □

Note that to obtain this estimate we actually split the Green's function $1/|x|$. To exploit this estimate along minimizing sequences we need to know that $h_M^r < 0$. It is here that we need the assumption (Φ3).

Lemma 3.5.

(a) *For every $M > 0$ we have $h_M^r < 0$.*

(b) *For every $0 < \bar{M} \leq M$ we have $h_{\bar{M}}^r \geq (\bar{M}/M)^{5/3} h_M^r$.*

Proof. For $\rho \in \mathcal{F}_M^r$ and $a, b > 0$ we define $\bar{\rho}(x) := a\rho(bx)$. Then

$$\begin{aligned} \int \bar{\rho}\,dx &= ab^{-3} \int \rho\,dx, \\ E_{\text{pot}}(\bar{\rho}) &= a^2 b^{-5} E_{\text{pot}}(\rho), \\ \int \Phi(\bar{\rho}) &= b^{-3} \int \Phi(a\rho)\,dx. \end{aligned}$$

To prove part (a) we fix a bounded and compactly supported function $\rho \in \mathcal{F}^r_M$ and choose $a = b^3$ so that $\bar\rho \in \mathcal{F}^r_M$ as well. By (Φ3) and since $3/n' > 1$,

$$\mathcal{H}^r_\mathcal{C}(\bar\rho) = b^{-3} \int \Phi(b^3\rho)\, dx + b\, E_{\text{pot}}(\rho) \le C\, b^{3/n'} + b\, E_{\text{pot}}(\rho) < 0, \qquad b \to 0,$$

and part (a) is established. As to part (b), we take $a = 1$ and $b = (M/\bar M)^{1/3} \ge 1$. For $\rho \in \mathcal{F}^r_M$ and $\bar\rho \in \mathcal{F}^r_{\bar M}$ rescaled with these parameters we find that

$$\begin{aligned}\mathcal{H}^r_\mathcal{C}(\bar\rho) &= b^{-3} \int \Phi(\rho)\, dx + b^{-5} E_{\text{pot}}(\rho)\\ &\ge b^{-5} \left(\int \Phi(\rho)\, dx + E_{\text{pot}}(\rho) \right) = \left(\frac{\bar M}{M}\right)^{5/3} \mathcal{H}^r_\mathcal{C}(\rho).\end{aligned} \tag{3.2}$$

Since for the present choice of a and b the map $\rho \mapsto \bar\rho$ is one-to-one and onto between $\mathcal{F}^r_M$ and $\mathcal{F}^r_{\bar M}$, this estimate proves part (b). ∎

COROLLARY 3.6. *Let $(\rho_i) \subset \mathcal{F}^r_M$ be a minimizing sequence of $\mathcal{H}^r_\mathcal{C}$. Then there exist $\delta_0 > 0$, $R_0 > 0$, $i_0 \in \mathbb{N}$, and a sequence of shift vectors $(a_i) \subset \mathbb{R}^3$ such that*

$$\int_{a_i + B_R} \rho_i(x)\, dx \ge \delta_0, \qquad i \ge i_0,\ R \ge R_0.$$

Proof. By Corollary 3.3, $(\|\rho_i\|_{1+1/n})$ is bounded. By Lemma 3.5(a) we have

$$E_{\text{pot}}(\rho_i) \le \mathcal{H}^r_\mathcal{C}(\rho_i) \le \frac{1}{2} h^r_M < 0, \qquad i \ge i_0,$$

for a suitable $i_0 \in \mathbb{N}$. Thus by Lemma 3.4 there exist $\delta_0 > 0$, $R_0 > 0$, and a sequence of shift vectors $(a_i) \subset \mathbb{R}^3$ as required. ∎

Finally, we will also need to exploit the well-known compactness properties of the solution operator of the Poisson equation.

LEMMA 3.7. *Let $(\rho_i) \subset L^{1+1/n}(\mathbb{R}^3)$ be bounded and*

$$\rho_i \rightharpoonup \rho_0 \text{ weakly in } L^{1+1/n}(\mathbb{R}^3).$$

(a) *For any $R > 0$,*

$$\nabla U_{\mathbf{1}_{B_R}\rho_i} \to \nabla U_{\mathbf{1}_{B_R}\rho_0} \text{ strongly in } L^2(\mathbb{R}^3).$$

(b) *If in addition (ρ_i) is bounded in $L^1(\mathbb{R}^3)$, $\rho_0 \in L^1(\mathbb{R}^3)$, and for any $\epsilon > 0$ there exist $R > 0$ and $i_0 \in \mathbb{N}$ such that*

$$\int_{|x| \ge R} |\rho_i(x)|\, dx < \epsilon, \qquad i \ge i_0,$$

then

$$\nabla U_{\rho_i} \to \nabla U_{\rho_0} \text{ strongly in } L^2(\mathbb{R}^3).$$

Proof. As to part (a), take any $R' > R$. Since $1 + 1/n > 4/3 > 6/5$, the mapping

$$L^{1+1/n}(\mathbb{R}^3) \ni \rho \mapsto \mathbf{1}_{B_{R'}} \nabla U_{\mathbf{1}_{B_R}\rho} \in L^2(B_{R'})$$

is compact. Thus the asserted strong convergence holds on $B_{R'}$. On the other hand,

$$\int_{|x|\geq R'} |\nabla U_{\mathbf{1}_{B_R}\rho_i}|^2 dx \leq \frac{C}{R'-R}\|\mathbf{1}_{B_R}\rho_i\|_1^2 \leq \frac{C}{R'-R}, \qquad i \in \mathbb{N}\cup\{0\},$$

which is arbitrarily small for R' large. As to part (b), we have for any $R > 0$,

$$\|\nabla U_{\rho_i} - \nabla U_{\rho_0}\|_2 \leq \|\nabla U_{\mathbf{1}_{B_R}\rho_i} - \nabla U_{\mathbf{1}_{B_R}\rho_0}\|_2 + \|\nabla U_{\mathbf{1}_{B_{R,\infty}}\rho_i} - \nabla U_{\mathbf{1}_{B_{R,\infty}}\rho_0}\|_2.$$

Using the extended Young's inequality, interpolation, and the boundedness of the sequence in $L^{1+1/n}(\mathbb{R}^3)$, we find that

$$\begin{aligned}\|\nabla U_{\mathbf{1}_{B_{R,\infty}}\rho_i} - \nabla U_{\mathbf{1}_{B_{R,\infty}}\rho_0}\|_2 &\leq C\left(\|\mathbf{1}_{B_{R,\infty}}\rho_i\|_{6/5} + \|\mathbf{1}_{B_{R,\infty}}\rho_0\|_{6/5}\right)\\ &\leq C\left(\|\mathbf{1}_{B_{R,\infty}}\rho_i\|_1^{(5-n)/6} + \|\mathbf{1}_{B_{R,\infty}}\rho_0\|_1^{(5-n)/6}\right).\end{aligned}$$

Given $\epsilon > 0$ we now choose $R > 0$ and $i_0 \in \mathbb{N}$ such that this is less than $\epsilon > 0$ for $i \geq i_0$, and recalling (a) completes the proof. □

We are now ready to prove the main result of this section.

Proof of Theorem 3.1. We split $\rho \in \mathcal{F}_M^r$ into three different parts:

$$\rho = \mathbf{1}_{B_{R_1}}\rho + \mathbf{1}_{B_{R_1,R_2}}\rho + \mathbf{1}_{B_{R_2,\infty}}\rho =: \rho_1 + \rho_2 + \rho_3;$$

the parameters $R_1 < R_2$ of the split are yet to be determined. With

$$I_{lm} := \iint \frac{\rho_l(x)\,\rho_m(y)}{|x-y|}, \qquad l,m = 1,2,3,$$

we have

$$\mathcal{H}_\mathcal{C}^r(\rho) = \mathcal{H}_\mathcal{C}^r(\rho_1) + \mathcal{H}_\mathcal{C}^r(\rho_2) + \mathcal{H}_\mathcal{C}^r(\rho_3) - I_{12} - I_{13} - I_{23}.$$

If we choose $R_2 > 2R_1$, then

$$I_{13} \leq \frac{C}{R_2}.$$

Next, we use the Cauchy–Schwarz inequality, the extended Young's inequality, and interpolation to get

$$\begin{aligned}I_{12} + I_{23} = \frac{1}{4\pi}\left|\int \nabla(U_1 + U_3)\cdot\nabla U_2 dx\right| &\leq C\|\rho_1 + \rho_3\|_{6/5}\|\nabla U_2\|_2\\ &\leq C\|\rho\|_{1+1/n}^{(n+1)/6}\,\|\nabla U_2\|_2.\end{aligned}$$

Using the estimates above and Lemma 3.5(b), we find with $M_l = \int \rho_l,\ l = 1,2,3,$

$$\begin{aligned}h_M^r - \mathcal{H}_\mathcal{C}^r(\rho) &\leq \left(1 - \left(\frac{M_1}{M}\right)^{5/3} - \left(\frac{M_2}{M}\right)^{5/3} - \left(\frac{M_3}{M}\right)^{5/3}\right) h_M^r\\ &\quad + C\left(R_2^{-1} + \|\rho\|_{1+1/n}^{(n+1)/6}\,\|\nabla U_2\|_2\right)\\ &\leq \frac{C}{M^2}\left(M_1M_2 + M_1M_3 + M_2M_3\right) h_M^r\\ &\quad + C\left(R_2^{-1} + \|\rho\|_{1+1/n}^{(n+1)/6}\,\|\nabla U_2\|_2\right)\\ &\leq Ch_M^r\, M_1\, M_3 + C\left(R_2^{-1} + \|\rho\|_{1+1/n}^{(n+1)/6}\,\|\nabla U_2\|_2\right);\end{aligned} \tag{3.3}$$

observe that by Lemma 3.5(a) $h^r_M < 0$ and that constants denoted by C are positive and depend on M and Φ but not on R_1 or R_2. We want to use (3.3) to show that up to a subsequence and a shift M_3 becomes small along any minimizing sequence for i large provided the splitting parameters are suitably chosen.

The sequence $T\rho_i := \rho_i(\cdot + a_i),\ i \in \mathbb{N}$, is minimizing and bounded in $L^{1+1/n}(\mathbb{R}^3)$ so there exists a subsequence, denoted by $(T\rho_i)$ again, such that $T\rho_i \rightharpoonup \rho_0$ weakly in $L^{1+1/n}(\mathbb{R}^3)$; cf. Corollary 3.3. Now choose $R_0 < R_1$ so that by Corollary 3.6, $M_{i,1} \geq \delta_0$ for i large. By (3.3),

$$(3.4)\ -C\, h^r_M \delta_0 M_{i,3} \leq \frac{C}{R_2} + C\,\|\nabla U_{0,2}\|_2 + C\|\nabla U_{i,2} - \nabla U_{0,2}\|_2 + \mathcal{H}^r_{\mathcal{C}}(T\rho_i) - h^r_M,$$

where $U_{i,l}$ is the potential induced by $\rho_{i,l}$ which in turn has mass $M_{i,l},\ i \in \mathbb{N}\cup\{0\}$, and the index $l = 1, 2, 3$ refers to the splitting. Given any $\epsilon > 0$ we increase $R_1 > R_0$ such that the second term on the right-hand side of (3.4) is small, say less than $\epsilon/4$. Next choose $R_2 > 2R_1$ such that the first term is small. Now that R_1 and R_2 are fixed, the third term in (3.4) converges to zero by Lemma 3.7(a). Since $(T\rho_i)$ is minimizing the remainder in (3.4) follows suit. Therefore, for i sufficiently large,

$$(3.5)\qquad \int_{a_i + B_{R_2}} T\rho_i = M - M_{i,3} \geq M - (-C\, h^r_M \delta_0)^{-1}\epsilon.$$

Clearly, $\rho_0 \geq 0$ a.e. By weak convergence we have that for any $\epsilon > 0$ there exists $R > 0$ such that

$$M \geq \int_{B_R} \rho_0\, dx \geq M - \epsilon,$$

which in particular implies that $\rho_0 \in L^1(\mathbb{R}^3)$ with $\int \rho_0 dx = M$. The functional $\rho \mapsto \int \Phi(\rho)\, dx$ is convex, so by Mazur's lemma and Fatou's lemma

$$\int \Phi(\rho_0)\, dx \leq \limsup_{i\to\infty} \int \Phi(T\rho_i)\, dx.$$

The strong convergence of the gravitational fields now follows by Lemma 3.7(b), and in particular,

$$\mathcal{H}^r_{\mathcal{C}}(\rho_0) \leq \limsup_{i\to\infty} \mathcal{H}^r_{\mathcal{C}}(\rho_i) = h^r_M$$

so that ρ_0 is a minimizer of $\mathcal{H}^r_{\mathcal{C}}$. □

4. Applications, symmetries, extensions. Although the main purpose of the present paper is to get a more general understanding of the techniques developed in [1, 2, 3, 4, 5, 11, 12] we want to at least indicate some possible applications of these techniques. First we should mention that [5] differs from the other papers in so far as there the Casimir functional is used as part of the constraint under which then the total energy is minimized. This made it possible to relax the growth conditions on Q—$0 < k \leq 7/2$ is covered in [5]—but since in the reduction process we turn $\mathcal{C}(f)+E_{\rm kin}(f)$ into a new functional of ρ, [5] seems to be outside the present framework.

We start with the observation, already noted in Theorem 2.1, that if $\rho_0 \in \mathcal{F}^r_M$ is a minimizer of $\mathcal{H}^r_{\mathcal{C}}$ with induced potential U_0, then

$$(4.1)\qquad \rho_0 = (\Phi')^{-1}_+(E_0 - U_0) := \begin{cases} (\Phi')^{-1}(E_0 - U_0), & U_0 < E_0,\\ 0, & U_0 \geq E_0,\end{cases}$$

and thus

$$\triangle U_0 = 4\pi (\Phi')_+^{-1}(E_0 - U_0) \tag{4.2}$$

on $\mathbb{R}^3$. The corresponding minimizer of $\mathcal{H}_\mathcal{C}$,

$$f_0 = \begin{cases} (Q')^{-1}(E_0 - E), & E < E_0, \\ 0, & E \geq E_0, \end{cases}$$

is a steady state of the Vlasov–Poisson system, since $E = E(x,v) = \frac{1}{2}|v|^2 + U_0(x)$ is a conserved quantity for the characteristics of the Vlasov equation with potential U_0 induced by $\rho_0 = \rho_{f_0}$. Steady states obtained in this manner have finite mass M, which is a necessary property for physically relevant steady states. We remark that the ansatz $f_0 = \phi(E_0 - E)$ reduces the stationary Vlasov–Poisson system to the semilinear Poisson equation

$$\triangle U_0 = 4\pi \int \phi\left(E_0 - \frac{1}{2}|v|^2 - U_0\right) dv,$$

which is exactly (4.2), provided Q can be chosen such that $(Q')^{-1} = \phi$ on $\mathbb{R}_+$ and $\phi = 0$ on $\mathbb{R}_-$. As far as the existence of steady states is concerned, our "reduced" approach allows us to cover $f_0 = (E_0 - E)_+^k$ with $-1 < k < 3/2$ which leads to (4.2) with right-hand side $C\,(E_0 - U_0)_+^n$ with $n = k + 3/2$ in the permissible range $]0, 3[$; note that the lower bound $k > -1$ is necessary to make the v-integral above converge. With the direct approach working with $\mathcal{H}_\mathcal{C}$ we were restricted to $0 < k < 3/2$.

The main feature of steady states obtained as minimizers in this manner is that their nonlinear stability. Since this is the main point in the investigations cited above, we do not go into this here. Instead, we briefly look a the role of symmetries in our problem. First we note that for any $\rho_0 \in \mathcal{F}_M^r$ its spherically symmetric decreasing rearrangement, denoted by ρ_0^*, also lies in $\mathcal{F}_M^r$ and satisfies

$$\int \Phi(\rho_0) = \int \Phi(\rho_0^*), \qquad E_{\mathrm{pot}}(\rho_0) \geq E_{\mathrm{pot}}(\rho_0^*)$$

with equality if and only if $\rho_0 = \rho_0^*(\cdot - x^*)$ for some $x^* \in \mathbb{R}^3$; cf. [8, Thms. 3.7, 3.9]. In particular, any minimizer of $\mathcal{H}_\mathcal{C}^r$ must be spherically symmetric with respect to some point in $\mathbb{R}^3$. If we are interested only in solving (4.2) or the stationary Vlasov–Poisson system, we therefore lose nothing if we restrict ourselves to the set of spherically symmetric functions in $\mathcal{F}_M^r$. The crucial part of the concentration-compactness argument simplifies considerably under this restriction.

Lemma 4.1. *Define*

$$R_0 = -\frac{3}{5}\frac{M^2}{h_M^r} > 0.$$

Let $\rho \in \mathcal{F}_M^r$ be spherically symmetric, $R > 0$, and

$$m := \int_{\{|x| \geq R\}} \rho.$$

Then the following estimate holds:

$$\mathcal{H}_\mathcal{C}^r(\rho) \geq h_M^r + \left[\frac{1}{R_0} - \frac{1}{R}\right](M - m)\,m.$$

If $R > R_0$, then for any spherically symmetric minimizing sequence $(\rho_i) \subset \mathcal{F}^r_M$ of $\mathcal{H}^r_\mathcal{C}$,

$$\lim_{i\to\infty} \int_{|x|\geq R} \rho_i = 0.$$

Proof. Clearly,

$$\mathcal{H}^r_\mathcal{C}(\rho) = \mathcal{H}^r_\mathcal{C}(\rho_1) + \mathcal{H}^r_\mathcal{C}(\rho_2) - \int \frac{\rho_1(x)\,\rho_2(y)}{|x-y|}dx\,dy,$$

where $\rho_1 = \mathbf{1}_{B_R}\rho$, $\rho_2 = \rho - \rho_1$. Due to spherical symmetry,

$$\begin{aligned}\int \frac{\rho_1(x)\,\rho_2(y)}{|x-y|}dx\,dy &= \frac{1}{4\pi}\int \nabla U_{\rho_1}\cdot\nabla U_{\rho_2}dx\\ &= \int_0^\infty \frac{4\pi}{r^2}\int_0^r \rho_1(s)\,s^2ds\,\frac{4\pi}{r^2}\int_0^r \rho_2(s)\,s^2ds\,r^2dr\\ &= \int_R^\infty \cdots dr \leq \frac{(M-m)\,m}{R}.\end{aligned}$$

Thus by Lemma 3.5,

$$\begin{aligned}\mathcal{H}^r_\mathcal{C}(\rho) &\geq h^r_{M-m} + h^r_m - \frac{(M-m)\,m}{R}\\ &\geq \left[\left(\frac{M-m}{M}\right)^{5/3} + \left(\frac{m}{M}\right)^{5/3}\right] h^r_M - \frac{(M-m)\,m}{R}\\ &\geq \left[1 - \frac{5}{3}\frac{M-m}{M}\frac{m}{M}\right] h^r_M - \frac{(M-m)\,m}{R},\end{aligned}$$

which is the first assertion of the lemma; note that the scaling transformations in the proof of Lemma 3.5 preserve spherical symmetry. Now take $R > R_0$ and assume that the second assertion is false so that up to a subsequence,

$$\lim_{i\to\infty} \int_{|x|\geq R} \rho_i = m > 0.$$

Choose $R_i > R$ such that

$$m_i := \int_{|x|\geq R_i} \rho_i = \frac{1}{2}\int_{|x|\geq R} \rho_i.$$

By the already established splitting estimate,

$$\mathcal{H}^r_\mathcal{C}(\rho_i) \geq h^r_M + \left[\frac{1}{R_0} - \frac{1}{R_i}\right](M-m_i)\,m_i \geq h^r_M + \left[\frac{1}{R_0} - \frac{1}{R}\right](M-m_i)\,m_i,$$

and with $i \to \infty$,

$$h^r_M \geq h^r_M + \left[\frac{1}{R_0} - \frac{1}{R}\right](M - m/2)\,m/2 > h^r_M,$$

a contradiction. □

The lemma above now replaces Lemma 3.4, Corollary 3.6, and the proof of (3.5), which relied on the fairly lengthy argument via (3.3) and (3.4). In addition, we get a somewhat sharper result on the minimizer:

$$\operatorname{supp}\rho_0 \subset B_{R_0}.$$

That spherical symmetry helps with compactness issues was already noted in [13]. The a priori restriction to the spherically symmetric case is undesirable in view of resulting stability assertions: These would then be restricted to spherically symmetric perturbations. Moreover, the symmetry simplification cannot be used if one does not a priori know that the minimizers will be spherically symmetric. One example for this situation is the construction of steady states with axial symmetry, say with respect to the x_3-axis, by making Q in addition depend explicitly on $x_1 v_2 - x_2 v_1$, the angular momentum with respect to the axis of symmetry. The same reduction procedure as before now gives a function Φ that depends in addition on $r = r(x) = \sqrt{x_1^2 + x_2^2}$, and minimizers will not be spherically symmetric. An investigation of axially symmetric steady states and their stability will be the content of [6].

To illustrate the method explained in section 3 let $\Phi = \Phi(r,\rho)$ satisfy the assumptions (Φ1)–(Φ3) uniformly in $r \geq 0$, and assume in addition that $\Phi(\cdot,\rho)$ is nonincreasing for every $\rho > 0$. Then all the arguments in section 3 go through the following: That Φ is nonincreasing in r is needed in the scaling argument (3.2)—note that $b \geq 1$ so $r(x)/b \leq r(x)$ there—and all the other estimates remain unchanged. Thus we obtain a minimizer ρ_0 with induced potential U_0. Let us take the specific example

$$\Phi(r,\rho) := \frac{2+r}{1+r}\,\rho^2, \qquad r \geq 0,\ \rho \geq 0.$$

Then the variational equation (4.1) reads as

$$\rho_0(x) = \frac{1}{2}\,\frac{1+r(x)}{2+r(x)}\,(E_0 - U_0(x))$$

on the support of ρ_0, and due to the explicit dependence on $r(x)$ this minimizer is not spherically symmetric, and the simpler arguments stated above for the spherically symmetric case do not apply. If one wishes to study the minimization of $\mathcal{H}_\mathcal{C}^r$ with $\Phi(\rho)$ generalized to $\Phi(x,\rho)$ the crucial step in the analysis which restricts the possible dependence on x is the scaling in Lemma 3.5.

Another situation where the minimizers will in general not be spherically symmetric arises if one includes in the Vlasov–Poisson system an exterior gravitational field, say $U_e = U_{\rho_e}$ with some fixed $\rho_e \in L^1_+ \cap L^{1+1/n}(\mathbb{R}^3)$. It is quite easy to check that all the analysis carried out in this paper extends to this case; only the potential energy needs to be modified accordingly:

$$\begin{aligned} E_{\mathrm{pot}}(\rho) &= -\frac{1}{2}\iint \frac{\rho(x)\,\rho(y)}{|x-y|}dx\,dy - \iint \frac{\rho(x)\,\rho_e(y)}{|x-y|}dx\,dy \\ &= \frac{1}{2}\int \rho(x) U_\rho(x)\,dx + \int \rho(x) U_e(x)\,dx, \end{aligned}$$

and if ρ_e is not spherically symmetric, then neither are possible minimizers.

Acknowledgment. This paper originates from my collaboration with Y. Guo, Brown University, whom I would like to thank for many stimulating discussions.

REFERENCES

[1] Y. Guo, *Variational method in polytropic galaxies*, Arch. Ration. Mech. Anal., 150 (1999), pp. 209–224.

[2] Y. Guo, *On the generalized Antonov's stability criterion*, Contemp. Math., 263 (2000), pp. 85–107.

[3] Y. Guo and G. Rein, *Stable steady states in stellar dynamics*, Arch. Ration. Mech. Anal., 147 (1999), pp. 225–243.

[4] Y. Guo and G. Rein, *Existence and stability of Camm type steady states in galactic dynamics*, Indiana Univ. Math. J., 48 (1999), pp. 1237–1255.

[5] Y. Guo and G. Rein, *Isotropic steady states in galactic dynamics*, Comm. Math. Phys., 219 (2001), pp. 607–629.

[6] Y. Guo and G. Rein, *Axially Symmetric Steady States in Galactic Dynamics*, in preparation.

[7] D. D. Holm, J. E. Marsden, T. Ratiu, and A. Weinstein, *Nonlinear stability of fluid and plasma equilibria*, Phys. Rep., 123 (1985), pp. 1–116.

[8] E. H. Lieb and M. Loss, *Analysis*, AMS, Providence, RI, 1996.

[9] P.-L. Lions, *The concentration-compactness principle in the calculus of variations. The locally compact case. Part* 1, Ann. Inst. H. Poincaré Anal. Non Linéaire, 1 (1984), pp. 109–145.

[10] J. Mawhin and M. Willem, *Critical Point Theory and Hamiltonian Systems*, Springer-Verlag, New York, 1989.

[11] G. Rein, *Flat steady states in stellar dynamics—existence and stability*, Comm. Math. Phys., 205 (1999), pp. 229–247.

[12] G. Rein, *Stability of spherically symmetric steady states in galactic dynamics against general perturbations*, Arch. Ration. Mech. Anal., to appear.

[13] W. A. Strauss, *Existence of solitary waves in higher dimensions*, Comm. Math. Phys., 55 (1977), pp. 149–162.

[14] G. Wolansky, *On nonlinear stability of polytropic galaxies*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 16 (1999), pp. 15–48.

SIAM J. MATH. ANAL.
Vol. 33, No. 4, pp. 913–929

STABILITY OF VISCOUS SHOCK WAVES FOR PROBLEMS WITH NONSYMMETRIC VISCOSITY MATRICES*

M. LIEFVENDAHL† AND G. KREISS†

Abstract. Sufficient conditions for nonlinear stability of viscous shock wave solutions of systems of conservation laws are given. The analysis applies to strong shocks of Lax type but is restricted to perturbations with zero mass. We use the Laplace transform and reduce the question of stability to a spectral condition for the resolvent equation of the linearized problem.

Key words. shock waves, nonlinear stability, conservation laws

AMS subject classifications. 35L65, 35B35, 35K55

PII. S0036141000374944

1. Introduction.

Traveling wave solutions occur in systems modeling physical phenomena such as gas dynamics, magnetohydrodynamics, and phase transitions. To be able to simulate such a problem it is important to understand the dynamic properties of the system. In this paper we consider the stability of viscous shock waves satisfying a parabolic system of conservation laws

$$v_t + f(v)_x = Bv_{xx},$$

where v approaches asymptotic states U_R and U_L as $x \to \pm\infty$. We assume that in the corresponding inviscid case the asymptotic states can be connected by a so-called Lax shock (see Lax [10]). In the case of sufficiently weak shocks, that is, when the difference between the asymptotic states is sufficiently small, existence and stability is understood. See, for instance, Kopell and Howard [7], Liu [13], Goodman [4], and Szepessy and Xin [17]. In the latter three papers nonlinear stability under increasingly more general perturbations is established. Under general conditions on the nonlinearity only stable viscous shock waves are possible.

For strong shocks, however, we do not expect sufficient conditions for stability to be as general as in the weak shock case. In Freistühler and Zumbrun [3] and in Liefvendahl and Kreiss [12] viscous shock waves of Lax type are constructed such that the linearized problem has solutions that grow exponentially. An example of a stability result is the paper by Matsumura and Nishihara [16], where a stability theorem for strong shock waves of a model system for compressible viscous gases is proven. The system is of mixed hyperbolic-parabolic type.

In this paper we assume the existence of a strong viscous shock wave of Lax type and prove a theorem stating sufficient conditions for stability under zero mass perturbations. The result is presented in greater detail in Liefvendahl [11]. It is an extension of the results in Kreiss and Kreiss [9], where the case $B = I$ is treated. In our approach the conditions for stability involve the spectral properties of the linearized problem. The theorem proved here is a special case of the results stated in Zumbrun and Howard [18]. The latter apply also to general mass perturbations and to nonconstant B.

*Received by the editors July 5, 2000; accepted for publication (in revised form) June 8, 2001; published electronically December 28, 2001.
http://www.siam.org/journals/sima/33-4/37494.html

†Department of Numerical Analysis and Computer Science, KTH, S-100 44 Stockholm, Sweden (mli@nada.kth.se, gunillak@nada.kth.se).

1.1. The main theorem. In this subsection we state the problem and all assumptions and formulate the main result. In the next subsection we outline the proof.

Consider the Cauchy problem for a system of n real conservation laws:

$$\begin{aligned} v_t + f(v)_x &= Bv_{xx}, \qquad t > 0, \quad x \in \mathbb{R}, \\ v(x,0) &= v_0(x). \end{aligned}$$

Here the matrix B is constant and has eigenvalues with positive real part and the function f is smooth.

We shall assume there is a traveling wave solution U, which without loss of generality can be assumed to be stationary. We make the following assumption concerning U.

Assumption 1. (i) There is a smooth solution $U(x)$ tending exponentially to constant states U_R, U_L as $x \to \pm\infty$; that is, there are constants $\beta > 0$ and K such that

$$\begin{aligned} |U(x) - U_R| &\le Ke^{-\beta x}, \\ |U(x) - U_L| &\le Ke^{\beta x}. \end{aligned}$$

(ii) The Jacobian matrix of f evaluated at U_R and U_L has real, distinct, nonzero eigenvalues. Let n_R denote the number of negative eigenvalues of $f'(U_R)$ and n_L the number of positive eigenvalues of $f'(U_L)$. Then $n_R + n_L = n + 1$.

(iii) Let $\alpha_i(\xi)$ denote the eigenvalues of the symbol $-i\xi A - \xi^2 B$, where $\xi \in \mathbb{R}$ and $A = A_R$ or A_L. Then there is a constant $\gamma > 0$ such that

$$\text{Re } \alpha_i(\xi) \le -\gamma\xi^2.$$

The last two conditions ensure that the shock wave is of Lax type and that the asymptotic states are stable; see Majda and Pego [15].

The zero mass of the perturbation enters in the following assumption on the initial data.

Assumption 2. Let the initial condition be of the form

$$v_0(x) = U(x) + \varepsilon \left(\tilde{v}_0(x)\right)_x, \quad |\varepsilon| << 1,$$

where $\tilde{v}_0$ and its derivatives have bounded L^1- and L^2-norms.

With $B = I$ the Assumptions 1 and 2 imply Assumption 1.1 in [9]. Also, note that our assumptions on f, A, and B are a special case of (H0)–(H4) in [18].

We shall use L^p-norms, for $p \in \{1, 2, \infty\}$, and the H^1-norm. For a function $f : \mathbb{R} \to \mathbb{R}^n$ with components $\{f_k\}_{k=1}^n$, these are defined by

$$\begin{aligned} \|f\|_{L^p(a,b)} &= \left(\sum_{k=1}^n \int_a^b |f_k(x)|^p dx\right)^{1/p}, \qquad p \in \{1,2\}, \\ \|f\|_{L^\infty(a,b)} &= \max_{1\le k\le n} \left(\text{ess sup}_{x\in(a,b)} |f_k(x)|\right), \\ \|f\|_{H^1(a,b)} &= \left(\|f\|^2_{L^2(a,b)} + \|f_x\|^2_{L^2(a,b)}\right)^{1/2}. \end{aligned}$$

When it is obvious which interval (a,b) to use or when it is the entire real axis we will not explicitly write it.

The aim of this paper is to supplement the above assumptions by structural conditions which imply nonlinear stability. We shall investigate stability by considering the equation for the perturbation from steady state. Since we will use the Laplace transform, we want to consider a problem with homogeneous initial data. Therefore we introduce

$$v(x,t) = U(x) + \varepsilon e^{-t}\left(\tilde{v}_0(x)\right)_x + \varepsilon u(x,t).$$

This leads to the following equation for the introduced function u:

$$u_t + (Au)_x + \varepsilon\left(Cu + g(u)\right)_x = Bu_{xx} - h_x, \tag{1}$$

with homogeneous initial data

$$u(x,0) = 0. \tag{2}$$

In (1) we introduced the following:

$$\begin{aligned} A(x) &= f'(U(x)),\\ C(x,t) &= f'(U + \epsilon e^{-t}\tilde{v}_{0x}) - f'(U),\\ h(x,t) &= \frac{1}{\epsilon}\left(f(U + \epsilon e^{-t}\tilde{v}_{0x}) - f(U)\right) - e^{-t}\left(\tilde{v}_0 + B\tilde{v}_{0xx}\right), \end{aligned}$$

and g, which is the rest term in the Taylor expansion of the nonlinearity, so it can be bounded according to $|g(u)| \le K|u|^2$.

Connected with (1) is the eigenvalue problem

$$B\varphi_{xx} - (A\varphi)_x = \mu\varphi, \quad \|\varphi\|_{L^2} < \infty. \tag{3}$$

Clearly zero is an eigenvalue with eigenfunction U_x. This eigensolution corresponds to the nonuniqueness of shock wave solutions (U can be shifted in space). A necessary condition for linear (and therefore also for nonlinear) stability is that there are no eigenvalues with positive real part.

Assumption 3. The only eigenvalue with $\operatorname{Re}\mu \ge 0$ is $\mu = 0$. The dimension of the corresponding eigenspace is 1.

The last assumption follows.

Assumption 4. The $n \times n$ matrix

$$M = \left(\, S_R^{II} \quad S_L^{I} \quad U_R - U_L \,\right)$$

is nonsingular. Here S_R^{II} consists of the eigenvectors of A_R corresponding to positive eigenvalues and S_L^{I} consists of the eigenvectors of A_L corresponding to negative eigenvalues.

The importance of the matrix M for shock wave stability has been known a long time. Assumption 4 appears as a condition for stability of viscous profiles in, for instance, [8], [9], [17], and [18]. It is shown to hold for all sufficiently weak shocks in [17] and to be necessary for stability in [18]. (Note the misprint concerning this in [18], first appearing on page 19; $\{r_j^{\pm} : a_j^{\pm} \lesseqgtr 0\}$ should be $\{r_j^{\pm} : a_j^{\pm} \gtreqless 0\}$.) For the inviscid problem the condition is necessary for stability; see Majda [14] and Erpenbeck [2]. This is also the situation for a corresponding discrete problem; see Bultelle, Grassin, and Serre [1].

We can now formulate our main result. We use the following concept of stability.

Definition 1. *Problem* (1), (2) *is nonlinearly stable under zero mass perturbations if the solution $u(x,t)$ remains smooth for all $t \geq 0$ and $||u(\cdot,t)||_{L^\infty}$ tends to zero for $t \to \infty$ for sufficiently small ϵ. In particular, we call the problem linearly stable if the convergence takes place for $\varepsilon = 0$.*

Our main theorem is as follows.

Theorem 1. *If Assumptions* 1–4 *are satisfied, then problems* (1)–(2) *are nonlinearly stable under zero mass perturbations.*

Note that we do not require the shock to be weak. Also, we give no temporal decay rate for the solution. This cannot be done without more severe requirements on the spatial decay of the initial perturbation.

1.2. Outline of the proof. Most parts of the proof of the theorem are identical with the proof of the corresponding theorem in [9]. In this paper we give details only of the parts that differ significantly, which of course concern the generalization $B \neq I$.

We will derive estimates for the corresponding linear problem. The linear problem is Laplace transformed in time, yielding the resolvent equation

$$B\hat{u}_{xx} - \Big(A(x)\hat{u}\Big)_x - s\hat{u} = \hat{h}_x, \quad \|\hat{u}(\cdot,s)\|_{L_2} < \infty. \tag{4}$$

As in [9], we want to prove the estimate

$$\|\hat{u}\|^2_{H^1} \leq K\left(\|\hat{h}\|^2_{L^2} + \|\hat{h}\|^2_{L^1}\right), \quad \operatorname{Re} s > 0. \tag{5}$$

From the estimate (5) nonlinear stability follows; see [9].

For sufficiently large $|s|$, (5) is obtained by integration by parts. This can be done since there is a constant matrix $H = H^*$ (a symmetrizer) which is positive definite and satisfies

$$\operatorname{Re}\ (\hat{u}_x, (HB + B^*H)\hat{u}_x) \geq \|\hat{u}_x\|^2_{L^2}.$$

Here $(\cdot,\cdot)$ denotes the complex L^2-inner product.

For $C \geq |s| \geq c > 0$ we need a result stating that for $A = A_R$ and A_L, all roots $\kappa(s)$ of the characteristic equation,

$$\det(sI - \kappa A - \kappa^2 B) = 0, \qquad \operatorname{Re}\, s > 0,$$

have real part bounded away from zero, and that there are precisely n roots with positive real part and n roots with negative real part. Then we can proceed as in [9], reducing the problem (4) to a bounded interval and deriving the estimate (5) by compactness arguments. The required result for the characteristic equation is given in section 2.1.

In [9] the treatment for the remaining part of the s plane relies heavily on the diagonalization of the constant coefficient problems obtained by letting $|x| \to \infty$ in the resolvent equation. In our case the method of proof must be modified, since only block-diagonalization is possible. The details of this part of the proof are given in this paper.

In section 2.1 we analyze the constant coefficient problems connected with the asymptotic states U_R and U_L. In the appendix we have collected some results for ODEs on a half-line that relate this analysis to the resolvent equation. In section 2.2 there is also a preliminary analysis of the resolvent equation when $s = 0$. In section 3 we use these results to prove the following result.

Lemma 1. *There exists constants K and $c > 0$ such that for $s \in \Omega_c := \{s \in \mathbb{C} : \operatorname{Re}\, s > 0, |s| \leq c\}$ the solution of* (4) *satisfies the estimate* (5).

2. Preliminary analysis. In the first part of this section we state properties of the limiting coefficients of the resolvent equation when $|x| \to \infty$. In the last part we analyze the resolvent equation for $s = 0$ in the case with forcing having bounded support.

2.1. Properties of the limiting coefficient matrices. The coefficients of the resolvent equation are nearly constant when $|x| \gg 1$. In this section we study the limiting coefficient matrices ($x \to \pm\infty$) when the resolvent equation is written as a first order system.

More precisely, we will study the eigenvalues and block-diagonalizing transformations for the matrix

$$D_0(s) := \begin{pmatrix} B^{-1}A_R & B^{-1} \\ sI & 0 \end{pmatrix}.$$

The treatment of the corresponding matrix when A_R is changed to A_L is completely analogous. Recall that A_R has distinct eigenvalues. Thus we can diagonalize it as

$$T_{1R}A_R S_{1R} = \begin{pmatrix} \Lambda^- & 0 \\ 0 & \Lambda^+ \end{pmatrix}.$$

Here the diagonal of Λ^- consists of $\lambda_1, \ldots, \lambda_{n_R}$, which are all negative. The diagonal of Λ^+ consists of $\lambda_{n_R+1}, \ldots, \lambda_n$, which are all positive.

We will now state a series of properties of the introduced matrices. Statements 1–4 below are established in [18] and play an equally fundamental role therein. Statement 5 is special to our approach and is new.

LEMMA 2. *Denote the eigenvalues of $B^{-1}A_R$ by $\{\tau_i\}_{i=1}^n$ and the eigenvalues of $D_0(s)$ by $\{\kappa_i(s)\}_{i=1}^{2n}$.*

1. *For $|s| \ll 1$ we have the expansions*

$$\kappa_i = \tau_i + \mathcal{O}(s^{1/n}), \qquad (6)$$

$$\kappa_{i+n} = -\frac{s}{\lambda_i} + \frac{s^2\tilde{b}_{ii}}{\lambda_i^3} + \mathcal{O}(s^3),$$

which hold for $i = 1, \ldots, n$. Here $\{\tilde{b}_{ii}\}_{i=1}^n$ denotes the diagonal elements of $T_{1R}BS_{1R}$.

2. *For $s \in \Omega := \{s \in \mathbb{C} : \mathrm{Re}\, s \geq 0, s \neq 0\}$, exactly n of the functions κ_i have positive real part and n have negative real part. We emphasize that for no $s \in \Omega$ is there a κ_i which is purely imaginary. Also, no κ_i crosses the imaginary axis when s varies in Ω.*

3. *The number of eigenvalues $\{\tau_i\}_{i=1}^n$ with positive real part is $n - n_R$ and the number with negative real part is n_R. We can block diagonalize $B^{-1}A_R$ according to*

$$T_{2R}B^{-1}A_R S_{2R} = \begin{pmatrix} A^+ & 0 \\ 0 & A^- \end{pmatrix},$$

where the τ_i with positive real part are eigenvalues of A^+ and those with negative real part are eigenvalues of A^-. We have $T_{2R} = S_{2R}^{-1}$.

4. $\mathrm{Re}\, \tilde{b}_{ii} > 0$ *for $i = 1, \ldots, n$.*

5. *In the region $\Omega_c := \{s \in \mathbb{C} : \text{Re } s > 0, |s| \leq c\}$, for a sufficiently small c, we have analytic matrices $S_R(s)$ and $T_R := S_R^{-1}$ such that*

$$T_R(s)D_0(s)S_R(s) = \begin{pmatrix} C^+(s) & 0 & 0 & 0 \\ 0 & C^-(s) & 0 & 0 \\ 0 & 0 & \mathcal{K}^+(s) & 0 \\ 0 & 0 & 0 & \mathcal{K}^-(s) \end{pmatrix}, \tag{7}$$

where

- *C^+ is an $(n-n_R)\times(n-n_R)$ matrix. The eigenvalues of C^+ are $\kappa_1, \ldots, \kappa_{n-n_R}$.*
- *C^- is an $n_R \times n_R$ matrix. The eigenvalues of C^- are $\kappa_{n-n_R+1}, \ldots \kappa_n$.*
- *$\mathcal{K}^+$ is an $n_R \times n_R$ diagonal matrix. The eigenvalues of $\mathcal{K}^+$ are $\kappa_{n+1}, \ldots, \kappa_{n+n_R}$.*
- *$\mathcal{K}^-$ is an $(n-n_R)\times(n-n_R)$ diagonal matrix. The eigenvalues of $\mathcal{K}^-$ are $\kappa_{n+n_R+1}, \ldots, \kappa_{2n}$.*
- *The block-diagonalizing transformations have the following form for $|s| \ll 1$:*

$$S_R(s) = \begin{pmatrix} S_{2R} & -A_R^{-1}S_{1R} \\ 0 & S_{1R} \end{pmatrix} + \mathcal{O}(s),$$
$$T_R(s) = \begin{pmatrix} T_{2R} & T_{2R}A_R^{-1} \\ 0 & T_{1R} \end{pmatrix} + \mathcal{O}(s). \tag{8}$$

Proof. Start with property 1. The κ_i are solutions of the characteristic equation $P(\kappa, s) := \det(\kappa^2 B - \kappa A_R - sI) = 0$. We see that P is a polynomial in κ and s, so the κ_i are algebraic functions. This means that the $\kappa_i(s)$ are functions which are analytic except at finitely many points; see [6, p. 119]. To determine the expansions we first solve $P(\kappa, 0) = 0$ and obtain the $\mathcal{O}(1)$ terms. For $\kappa_{n+1}, \ldots, \kappa_{2n}$ this term is zero. For $\kappa_1, \ldots, \kappa_n$ we do not need more information; the remainder is in the worst case $\mathcal{O}(s^{1/n})$. For $\kappa_{n+1}, \ldots, \kappa_{2n}$ we continue by grouping terms of $P(\gamma_i s, s)$ according to powers in s. We will determine $\{\gamma_i\}_{i=1}^n$; here γ_i correspond to κ_{i+n}. We neglect higher order terms and set the expression to zero, which gives γ_i. To obtain the $\mathcal{O}(s^2)$ terms we use the ansatz $\kappa_{i+n} = -s/\lambda_i + s^2\theta_i$ and determine θ_i. The details are given in [11].

To prove property 2, we first study the characteristic equation for $|s| \gg 1$,

$$\det(\kappa^2 B - sI) \approx 0,$$

which yields $\kappa_i \approx \pm\sqrt{s/\beta_i}$, where β_i are the eigenvalues of B. This means that for $|s| \gg 1$ and $s \in \Omega$ half of the κ_i have positive real part and half of them have negative real part. Now we assume that one of the κ_i crosses the imaginary axis when s varies in Ω. Then

$$\det(-\xi_0^2 B - i\xi_0 A_R - s_0 I) = 0 \tag{9}$$

for some $\xi_0 \in \mathbb{R}$. According to (9), s_0 is an eigenvalue of the symbol $-\xi_0^2 B - i\xi_0 A_R$. Assumption 1 then implies Re $s_0 < -\gamma\xi_0^2$, which contradicts $s_0 \in \Omega$. Thus the number of κ_i with positive and negative real part, respectively, is n for all $s \in \Omega$.

Property 3 follows by using the expansions of κ_i with a real s and counting the number of eigenvalues in the right and left complex half planes.

Property 4 follows from the expansion for $\kappa_{n+1}, \dots, \kappa_{2n}$ for a purely imaginary s and the requirement that no κ_i can be purely imaginary.

To prove property 5, we note that the information contained in the expansions in (6), concerning grouping of eigenvalues of with positive and negative real part, and the order of the branching of the $\mathcal{O}(s)$ eigenvalues, make Theorem 8, page 70, in [5] applicable. This theorem is stated in terms of eigenprojections corresponding to eigenvalues. Restating it in terms of block diagonal transformations we get the existence and analyticity of S_R and T_R. The expansions of the matrices S_R and T_R then follow by inspection; see [11] for details. This concludes the proof. □

Finally we introduce notation for submatrices of the block-diagonalizing transformations introduced in this section. We partition the matrices T_{1R} and T_{2R} according to

$$T_{1R} = \begin{pmatrix} T_{1R}^{I} \\ T_{1R}^{II} \end{pmatrix}, \qquad T_{2R} = \begin{pmatrix} T_{2R}^{I} \\ T_{2R}^{II} \end{pmatrix},$$

where T_{1R}^{I} contains the first n_R rows of T_{1R}, and T_{1R}^{II} contains the last $n - n_R$ rows. Correspondingly, T_{2R}^{I} contains the first $n - n_R$ rows of T_{2R}, and T_{2R}^{II} the last n_R rows.

We partition the matrix T_R into submatrices corresponding to the blocks on the diagonal of the matrix in the right-hand side of (7),

$$T_R(s) = \begin{pmatrix} T_R^{I}(s) \\ T_R^{II}(s) \\ T_R^{III}(s) \\ T_R^{IV}(s) \end{pmatrix}.$$

The partition is done so that T_R^I has $(n - n_R)$ rows, T_R^{II} has n_R rows, etc.

2.2. Solution at the zero eigenvalue when the forcing has bounded support. The resolvent equation (4), with $s = 0$ and forcing stemming from initial data with zero mass, can be integrated once. We then obtain the problem

$$u_x = B^{-1}Au + f, \tag{10}$$

where we have dropped the ˆ-notation for the Laplace-transformed function. The forcing f is not directly related to $\hat{h}$. Instead we assume supp $f \subset [-l_0, l_0]$ for some l_0 and $f \in L^1 \cap L^2$. Studying (10) on an interval $[l, \infty)$ with $l > l_0$, we can apply Lemma 4. According to the lemma there exists an l_1, which we take larger than l_0, such that for $l > l_1$ the following holds. For solutions of the ODE (10), the requirement that u is bounded is equivalent with the condition

$$R(l)u(l) = 0, \tag{11}$$

where $R(l)$ can be written as

$$R(l) = T_{2R}^{I} - e^{-\beta l} P(l) T_{2R}^{II}. \tag{12}$$

Here T_{2R}^{I} and T_{2R}^{II} were introduced in the preceding section and P satisfies $|P(l)| \leq K$.

The argument above can also be applied to (10) on an interval $(-\infty, -l]$ with $l > l_0$ and we obtain boundary conditions

$$R(-l)u(-l) = 0, \tag{13}$$

where we, in the same manner, have the expression $R(-l) = T_{2L}^{II} - e^{-\beta l} P(-l) T_{2L}^{I}$.

Now we consider (10) for $x \in [-l, l]$ with the boundary conditions (11) and (13). This problem consists of n first order ODEs together with $n-1$ boundary conditions. We also note that any solution of this problem is also a solution of (10) on the whole real axis, and vice versa. The reason for replacing the ODE on the real axis with that on a bounded interval is to make it possible to estimate the solutions via a "compactness" argument, which we describe below.

Next we derive a representation formula for the solution by introducing a fundamental matrix $\Psi(x)$ for the homogeneous version of (10). The n columns of Ψ are linearly independent solutions of (10) with $f = 0$. The solution of the inhomogeneous problem can now be written as

$$u(x) = \Psi(x) \left[u(0) + \int_0^x \Psi^{-1}(\xi) f(\xi) d\xi \right].$$

Inserting this into the boundary conditions at $x = \pm l$ we obtain the following system of $n - 1$ linear equations for the components of $u(0)$:

$$\text{(14)} \qquad \begin{pmatrix} R(l)\Psi(l) \\ R(-l)\Psi(-l) \end{pmatrix} u(0) = \begin{pmatrix} R(l)\Psi(l) \int_0^l \Psi^{-1}(\xi) f(\xi) d\xi \\ -R(-l)\Psi(-l) \int_{-l}^0 \Psi^{-1}(\xi) f(\xi) d\xi \end{pmatrix}.$$

For the homogeneous problem, $f = 0$, we have zero in the right-hand side of (14). According to Assumption 3, the solutions of the homogeneous problem have the form $u = \alpha\varphi_0$ for an arbitrary α, i.e., the eigenspace of zero has dimension one. This means that the solutions of (14) in the homogeneous case are $u(0) = \alpha\varphi_0(0)$.

The solutions of (14) are not unique. Our goal is to construct one solution which can be estimated in terms of the forcing function f. This is done by adding one more condition, which is that we choose the solution $u(0)$ of (14) that has minimal Euclidean length.

The above procedure gives a function $u(x)$, which can be written as

$$u(x) = \alpha\varphi_0 + \Psi(x) \left[u_0 + \int_0^x \Psi^{-1}(\xi) f(\xi) d\xi \right].$$

Here u_0 is any solution of (14) and α is determined by the requirement that $\alpha\varphi_0(0) + u_0$ be the minimal solution of (14). We have the estimate

$$|u(0)| \le K \left| \int_{-l}^{l} \Psi^{-1}(\xi) f(\xi) d\xi \right| \le K \max_{x \in [-l,l]} \left| \Psi^{-1}(x) \right| \|f\|_{L^1}.$$

We obtain the last inequality since $|\Psi^{-1}(x)|$ is a continuous function on a compact domain, and therefore is bounded. The same argument holds for $|\Psi(x)|$. We also need to use

$$\|f\|_{L^1[-l,l]} \le K\sqrt{l} \|f\|_{L^2[-l,l]}.$$

Using the solution formula it is now easy to derive

$$\text{(15)} \qquad \|u\|_{L^1[-l,l]} + \|u\|_{L^2[-l,l]} + \|u\|_{L^\infty[-l,l]} \le K\|f\|_{L^2}.$$

To extend the solution from $[-l, l]$ to the entire real axis we use Lemma 3 with the now-determined boundary values. This gives the estimate

$$\text{(16)} \qquad \|u\|_{L^1[l,\infty)} + \|u\|_{L^2[l,\infty)} + \|u\|_{L^\infty[l,\infty)} \le K|u(l)|$$

and the corresponding one to the left. Since the right-hand side in (16) can be estimated by the L^∞-norm of u, we can use (15) to get an estimate in terms of f. Combining the estimates and using the triangle inequality in (10) we get

$$\|u\|_{H^1} + \|u\|_{L^1} + \|u\|_{L^\infty} \le K\|f\|_{L^2}.$$

3. The resolvent estimate near the zero eigenvalue. In this section we prove Lemma 1. The proof proceeds by splitting the unknown into several terms which satisfies simpler problems. These auxiliary problems are treated in sections 3.1–3.3, then we complete the proof by summing the constructions in section 3.4.

3.1. Reduction of the forcing to $\mathcal{O}(s)$. In our procedure to solve and estimate the solution of (4), we start by splitting the unknown $\hat{u} = u_1 + u_2$, where u_1 solves the equation

$$u_{1x} = B^{-1}Au_1 + \hat{h}. \tag{17}$$

When u_1 has been determined, u_2 must satisfy the following equation, which is derived by inserting $\hat{u} = u_1 + u_2$ into (4):

$$su_2 + (Au_2)_x = Bu_{2xx} - su_1. \tag{18}$$

Here we see that u_1, multiplied by the small s, occurs as forcing. Equation (18) is treated in section 3.2. To solve (17) we split the unknown further:

$$u_1 = u_{1M} + \varphi_R u_{1R} + \varphi_L u_{1L}.$$

Here φ_R and φ_L are smooth monotone cut-off functions,

$$\varphi_R = \begin{cases} 1, & x > l, \\ 0, & x < l-1, \end{cases} \qquad \varphi_L = \begin{cases} 1, & x < -l, \\ 0, & x > -l+1, \end{cases}$$

where l will be given below. The function u_{1R} solves the problem

$$\begin{aligned} &u_{1Rx} = B^{-1}Au_{1R} + \hat{h}, \qquad && x \in [l-1, \infty), \\ &T^I_{2R} u_{1R}(l-1) = 0, && \end{aligned} \tag{19}$$

and u_{1L} solves the corresponding problem on the interval $(-\infty, -l+1]$ with T^{II}_{2L} as coefficient matrix in the linear homogeneous boundary conditions at $x = -l+1$.

The remaining part of u_1 must satisfy

$$u_{1Mx} = B^{-1}Au_{1M} + h_M, \tag{20}$$

where $h_M = (1-\varphi_R-\varphi_L)\hat{h} - \varphi_{Rx}u_{1R} - \varphi_{Lx}u_{1L}$. We note that $\operatorname{supp} h_M \subset [-l, l]$. The equation for u_{1M} will be considered below after we have determined u_{1R} and u_{1L}.

If l is chosen sufficiently large, the problems for u_{1R} and u_{1L} are of the type treated in Lemma 3. We thus have unique solutions u_{1R} and u_{1L} for the problems to the right and left, respectively. The function u_{1R} satisfies the estimate

$$\|u_{1R}\|_{L^1[l-1,\infty)} + \|u_{1R}\|_{H^1[l-1,\infty)} + \|u_{1R}\|_{L^\infty[l-1,\infty)} \le K\left(\|\hat{h}\|_{L^1} + \|\hat{h}\|_{L^2}\right). \tag{21}$$

The function u_{1L} satisfies the corresponding estimate, i.e., with all R-subscripts replaced by L-subscripts and the interval changed from $[l-1,\infty)$ to $(-\infty, -l+1]$.

Now we study the equation for u_{1M}. Combining the expression for h_M with the estimate of u_{1R} and u_{1L} in terms of $\hat{h}$, we obtain an estimate of h_M in terms of $\hat{h}$:

$$\|h_M\|_{L^1} + \|h_M\|_{L^2} \leq K\left(\|\hat{h}\|_{L^1} + \|\hat{h}\|_{L^2}\right).$$

Equation (20) is of the type treated in section 2.2. Using the construction described there, we obtain a function u_{1M}, which solves (20) and can be estimated in terms of h_M and thus also in terms of $\hat{h}$:

$$\|u_{1M}\|_{H^1} + \|u_{1M}\|_{L^1} \leq K\left(\|\hat{h}\|_{L^1} + \|\hat{h}\|_{L^2}\right).$$

The last step in this section is to combine the constructions and estimates of u_{1R}, u_{1L}, and u_{1M}. This yields the function u_1 and the following estimate for it:

$$\|u_1\|_{H^1} + \|u_1\|_{L^1} \leq K\left(\|\hat{h}\|_{L^1} + \|\hat{h}\|_{L^2}\right). \tag{22}$$

3.2. Cut-off of forcing to obtain bounded support. In this section we will study (18), where u_1 has the role of forcing. We rewrite the equation to a system of first order equations by introducing the variable $v_2 = Bu_{2x} - Au_2$. This leads to

$$\begin{pmatrix} u_2 \\ v_2 \end{pmatrix}_x = \begin{pmatrix} B^{-1}A & B^{-1} \\ sI & 0 \end{pmatrix}\begin{pmatrix} u_2 \\ v_2 \end{pmatrix} + s\begin{pmatrix} 0 \\ u_1 \end{pmatrix}.$$

To solve this problem we split the unknown:

$$\begin{pmatrix} u_2 \\ v_2 \end{pmatrix} = \begin{pmatrix} u_{2M} \\ v_{2M} \end{pmatrix} + \varphi_R\begin{pmatrix} u_{2R} \\ v_{2R} \end{pmatrix} + \varphi_L\begin{pmatrix} u_{2L} \\ v_{2L} \end{pmatrix}.$$

Here φ_R and φ_L are, as before, smooth and monotone cut-off functions for the right and left half-infinite intervals, respectively. The location where the cut-off is performed in this section is chosen independently of the cut-off location in section 3.1. For u_{2R} and v_{2R} we have the problem

$$\begin{pmatrix} u_{2R} \\ v_{2R} \end{pmatrix}_x = \begin{pmatrix} B^{-1}A & B^{-1} \\ sI & 0 \end{pmatrix}\begin{pmatrix} u_{2R} \\ v_{2R} \end{pmatrix} + s\begin{pmatrix} 0 \\ u_1 \end{pmatrix}, \qquad x \in [l-1, \infty), \tag{23}$$

where l remains to be fixed. We choose homogeneous boundary conditions

$$\begin{pmatrix} T_R^I(s) \\ T_R^{III}(s) \end{pmatrix}\begin{pmatrix} u_{2R}(l-1) \\ v_{2R}(l-1) \end{pmatrix} = 0,$$

where T_R^I and T_R^{III} were introduced in section 2.1. For u_{2L} and v_{2L} we have the corresponding problem on the interval $(-\infty, -l+1]$, with T_L^{II} and T_L^{IV} as submatrices in the homogeneous boundary conditions at $x = -l+1$.

The remaining parts of u_2 and v_2, u_{2M} and v_{2M}, must satisfy the equation

$$\begin{pmatrix} u_{2M} \\ v_{2M} \end{pmatrix}_x = \begin{pmatrix} B^{-1}A & B^{-1} \\ sI & 0 \end{pmatrix}\begin{pmatrix} u_{2M} \\ v_{2M} \end{pmatrix} + s\begin{pmatrix} g \\ h \end{pmatrix}, \tag{24}$$

where

$$\begin{cases} sg = -\varphi_{Rx}u_{2R} - \varphi_{Lx}u_{2L}, \\ sh = s(1-\varphi_R-\varphi_L)u_1 - \varphi_{Rx}v_{2R} - \varphi_{Lx}v_{2L}. \end{cases} \tag{25}$$

Equation (24) will be studied in section 3.3.

If l is chosen sufficiently large, the problems for $(u_{2R}, v_{2R})^T$ and $(u_{2L}, v_{2L})^T$ are of the type treated in Lemma 5. We thus have unique, bounded solutions of both problems. The functions u_{2R} and v_{2R} satisfy the estimates

$$\begin{aligned} &\|u_{2R}(\cdot,s)\|_{H^1[l-1,\infty)} + \|v_{2R}(\cdot,s)\|_{H^1[l-1,\infty)} \leq K\left(\|u_1\|_{L^1} + \|u_1\|_{L^2}\right), \\ &\|u_{2R}(\cdot,s)\|_{L^\infty[l-1,\infty)} + \|v_{2R}(\cdot,s)\|_{L^\infty[l-1,\infty)} \leq |s|K\left(\|u_1\|_{L^1} + \|u_1\|_{L^2}\right). \end{aligned} \tag{26}$$

Here the L^2 estimate follows from the lemma. The L^2 estimate of the derivative, and thereby the H^1 estimate, follows from the triangle inequality applied to (23). For the functions u_{2L} and v_{2L} we have the corresponding estimates on the interval $(-\infty, -l+1]$. We also observe that, using (22), we obtain an estimate in terms of the original forcing $\hat{h}$.

3.3. Approximation of the component in the zero eigenspace. We now study (24). Recall that the forcing is $\mathcal{O}(s)$ and has support in a bounded interval, for which we introduce the notation $[-l_0, l_0]$. The forcing functions g and h can, via (22), (25), and (26), be estimated in terms of the original forcing $\hat{h}$. For $|s| \ll 1$ the problem is nearly singular, and we expect the solution to have a large component in the zero eigenspace of (3). This motivates the following splitting of u_{2M}:

$$u_{2M} = u_3 + \alpha\tilde{\varphi}_0 + u_4, \tag{27}$$

where

$$\tilde{\varphi}_0(x) = \begin{cases} \varphi_0(x), & |x| \leq l, \\ 0, & |x| \geq l, \end{cases}$$

and l will be determined below. We introduce $v_3 := v_{2M}$ for notational convenience. We use $\tilde{\varphi}_0$ instead of φ_0 in the ansatz (27) because we want to retain the bounded support of the forcing in the equations for u_3 and v_3, which are

$$\begin{pmatrix} u_3 \\ v_3 \end{pmatrix}_x = \begin{pmatrix} B^{-1}A & B^{-1} \\ sI & 0 \end{pmatrix} \begin{pmatrix} u_3 \\ v_3 \end{pmatrix} + s\begin{pmatrix} g \\ h - \alpha\tilde{\varphi}_0 \end{pmatrix}. \tag{28}$$

We can apply Lemma 6 to (28) on an interval $[l, \infty)$. According to the lemma there exists an l_1, which we choose larger than l_0, such that for $l > l_1$ the following holds. For solutions of the ODE (28), the requirement that the solution is bounded is equivalent to the condition

$$\tilde{R}(l,s)\begin{pmatrix} u_3(l) \\ v_3(l) \end{pmatrix} = 0, \tag{29}$$

where $\tilde{R}$ is described in the lemma. Lemma 6 can also be applied to (24) on the interval $(-\infty, -l+1]$ and we obtain boundary conditions

$$\tilde{R}(-l,s)\begin{pmatrix} u_3(-l) \\ v_3(-l) \end{pmatrix} = 0. \tag{30}$$

We use an iteration to construct α, u_3, and v_3 on $[-l, l]$, so that estimates in terms of g and h are possible, and (28), (29), and (30) are satisfied. When u_3, v_3, and α have been determined we will, in section 3.4, consider the equation for u_4, which we introduced in (27). The iteration, which is described in detail below, is an L^∞-contraction for sufficiently small $|s|$. The limit is our solution u_3, v_3, and α.

The first iterate satisfies the problem

$$\begin{aligned}&\begin{pmatrix} u^{(1)} \\ v^{(1)} \end{pmatrix}_x = \begin{pmatrix} B^{-1}A & B^{-1} \\ 0 & 0 \end{pmatrix}\begin{pmatrix} u^{(1)} \\ v^{(1)} \end{pmatrix} + s\begin{pmatrix} g \\ h-\alpha^{(1)}\varphi_0 \end{pmatrix},\\ &\tilde{R}(l,0)\begin{pmatrix} u^{(1)}(l) \\ v^{(1)}(l) \end{pmatrix} = 0, \qquad \tilde{R}(-l,0)\begin{pmatrix} u^{(1)}(-l) \\ v^{(1)}(-l) \end{pmatrix} = 0.\end{aligned} \tag{31}$$

The nth iterate satisfies the problem

$$\begin{aligned}&\begin{pmatrix} u^{(n)} \\ v^{(n)} \end{pmatrix}_x = \begin{pmatrix} B^{-1}A & B^{-1} \\ 0 & 0 \end{pmatrix}\begin{pmatrix} u^{(n)} \\ v^{(n)} \end{pmatrix} + s\begin{pmatrix} 0 \\ u^{(n-1)}-\alpha^{(n)}\varphi_0 \end{pmatrix},\\ &\tilde{R}(l,0)\begin{pmatrix} u^{(n)}(l) \\ v^{(n)}(l) \end{pmatrix} = (\tilde{R}(l,s)-\tilde{R}(l,0))\begin{pmatrix} u^{(n-1)}(l) \\ v^{(n-1)}(l) \end{pmatrix},\\ &\tilde{R}(-l,0)\begin{pmatrix} u^{(n)}(-l) \\ v^{(n)}(-l) \end{pmatrix} = (\tilde{R}(-l,s)-\tilde{R}(-l,0))\begin{pmatrix} u^{(n-1)}(-l) \\ v^{(n-1)}(-l) \end{pmatrix}.\end{aligned} \tag{32}$$

First we determine $v^{(1)}$, $\alpha^{(1)}$. From (31), using the expression (47) for $\tilde{R}(l,0)$, we extract the following equations for $v^{(1)}$, $\alpha^{(1)}$:

$$v_x^{(1)} = sh - s\alpha^{(1)}\varphi_0, \qquad T_{1R}^{I}v^{(1)}(l) = 0, \qquad T_{1L}^{II}v^{(1)}(-l) = 0. \tag{33}$$

The last two expressions imply that we can write $v(l) = S_R^{II}c^{II}$ and $v(-l) = S_L^{I}c^{I}$, where S_R^{II} and S_L^{I} were introduced in Assumption 4. Here the components of c^I and c^{II} together comprise $n-1$ unknowns. Using these expressions and integrating the first equation in (33) we arrive at the following $n \times n$ system of linear equations:

$$\left(S_{1L}^{I};\, S_{1R}^{II};\, \int_{-l}^{l}\varphi_0\,dx\right)\begin{pmatrix} c^I \\ c^{II} \\ s\alpha^{(1)} \end{pmatrix} = s\int_{-l}^{l} h\,dx. \tag{34}$$

We have

$$\left|U_R - U_L - \int_{-l}^{l}\varphi_0\,dx\right| \le Ke^{-\beta l}$$

and thus, by Assumption 4, (34) is solvable for sufficiently large l. Since $v^{(1)}$ is expressed by an integral of h, it is easy to derive

$$\begin{aligned}\|v^{(1)}\|_{L^\infty[-l,l]} + \|v^{(1)}\|_{H^1[-l,l]} &\le |s|K\|h\|_{L^2},\\ |\alpha^{(1)}| &\le K\|h\|_{L^2}.\end{aligned} \tag{35}$$

Next we turn to the problem for $u^{(1)}$. The boundary conditions will be inhomogeneous; e.g., to the right, we have

$$R(l)u^{(1)}(l) = -\tilde{R}_{12}(l)v^{(1)}(l).$$

The standard procedure to make the boundary conditions homogeneous is to make an ansatz $u^{(1)} = w + \tilde{u}$ and choose w as a smoothly varying function which satisfies the boundary conditions. The problem for $\tilde{u}$ is

$$\begin{aligned}&\tilde{u}_x = B^{-1}A\tilde{u} + sg^{(1)},\\ &R(l)\tilde{u}(l) = 0, \qquad R(-l)\tilde{u}(-l) = 0,\end{aligned}$$

where $sg^{(1)} = sg - w_x + B^{-1}Aw$. The L^2-norm of the function $g^{(1)}$ can be estimated in terms of g and h via (35). The problem for $\tilde{u}$ is of the type treated in section 2.2, and using the construction described there, we obtain a function $\tilde{u}$ which solves the above problem. Using $u^{(1)} = w + \tilde{u}$ we get a solution of (31), which satisfies the estimate

$$\|u^{(1)}\|_{L^\infty[-l,l]} + \|u^{(1)}\|_{H^1[-l,l]} \le |s|K\left(\|g\|_{L^2} + \|h\|_{L^2}\right).$$

This completes the construction and estimate of the first iterate. The treatment of the nth iterate is similar, only the forcing functions in the equations are changed. First we determine $v^{(n)}$ and $\alpha^{(n)}$. The boundary conditions are now inhomogeneous. From (32) and (47) we have

$$T^I_{1R} v^{(n)}(l) = sQ_1(l,s)\begin{pmatrix} u^{(n-1)}(l) \\ v^{(n-1)}(l) \end{pmatrix}.$$

Thus

$$v^{(n)}(l) = S^{II}_{1R}c^{II} + b^{(n)}(l), \tag{36}$$

where

$$b^{(n)}(l) = sS^I_{1R}Q_1\begin{pmatrix} u^{(n-1)}(l) \\ v^{(n-1)}(l) \end{pmatrix}.$$

In the same manner we obtain the following expression to the left:

$$v^{(n)}(-l) = S^I_{1L}c^I + b^{(n)}(-l). \tag{37}$$

Thus we can express $v^{(n)}(l)$ and $v^{(n)}(-l)$ in terms of the $n-1$ unknowns, c^I and c^{II}. Next we integrate the equation for $v^{(n)}$ and use the expressions (36) and (37) to obtain

$$\left(S^I_{1L};\, S^{II}_{1R};\, \int_{-l}^{l}\varphi_0\,dx\right)\begin{pmatrix} c^I \\ c^{II} \\ s\alpha^{(n)} \end{pmatrix} = s\int_{-l}^{l} u^{(n-1)}\,dx + b^{(n)}(-l) + b^{(n)}(-l).$$

As before, the system can be solved for sufficiently large l, and the solution can be estimated as

$$\begin{aligned} &\|v^{(n)}\|_{L^\infty[-l,l]} + \|v^{(n)}\|_{H^1[-l,l]} + |s\alpha^{(n)}| \\ &\quad\le sK\left(\|u^{(n-1)}\|_{L^2[-l,l]} + \|u^{(n-1)}\|_{L^\infty[-l,l]} + \|v^{(n-1)}\|_{L^\infty[-l,l]}\right). \end{aligned} \tag{38}$$

The problem for $u^{(n)}$ also has inhomogeneous boundary conditions. They are treated with the standard procedure described above. Referring to the construction in section 2.2, we obtain a solution which satisfies

$$\begin{aligned} &\|v^{(n)}\|_{L^\infty[-l,l]} + \|v^{(n)}\|_{H^1[-l,l]} \\ &\le sK\left(\|u^{(n-1)}\|_{L^2[-l,l]} + \|v^{(n-1)}\|_{L^2[-l,l]} + \|u^{(n-1)}\|_{L^\infty[-l,l]} + \|v^{(n-1)}\|_{L^\infty[-l,l]}\right). \end{aligned} \tag{39}$$

The construction is now complete. We see that for sufficiently small $|s|$ the iteration defines a contraction. The solution of our original problem (28), (29), and (30) is given by the uniformly convergent sums

$$u_3 = \sum_{n=1}^{\infty} u^{(n)}, \qquad v_3 = \sum_{n=1}^{\infty} v^{(n)}, \qquad \alpha = \sum_{n=1}^{\infty} \alpha^{(n)}.$$

Summing the estimates for the iterates, we obtain

$$\begin{aligned} \|u_3\|_{H^1[-l,l]} &\le |s| K \left(\|h\|_{L^2} + \|g\|_{L^2} \right), \\ |\alpha| &\le K \left(\|h\|_{L^2} + \|g\|_{L^2} \right). \end{aligned} \tag{40}$$

3.4. Summary of the splittings. Before the summing of the splitting of $\hat{u}$ into terms, we extend the function u_3 to the intervals $[l, \infty)$ and $(-\infty, -l]$. Since (28) is homogeneous in these regions, we can apply Lemma 5. This gives estimates that together with (40) yield

$$\|u_3\|_{H^1} \le K \left(\|h\|_{L^2} + \|g\|_{L^2} \right).$$

Summing the splittings from the sections above, we have the following expression of the original unknown in terms of the introduced parts:

$$\hat{u} = u_1 + \varphi_R u_{2R} + \varphi_L u_{2L} + u_3 + \alpha \varphi_0 + u_4.$$

Here the function u_4 satisfies the equation

$$s u_4 + (A u_4)_x = B u_{4xx} + s\alpha(\varphi_0 - \tilde{\varphi}_0).$$

This means that the function u_4 satisfies the same type of problem as u_2; see (18). The difference is that the norms of the forcing has been reduced by a factor $\mathcal{O}(e^{-\beta l})$. This leads us to define an iteration to determine $\hat{u}$. The first iterate $\hat{u}^{(1)}$ is given by

$$\hat{u}^{(1)} = u_1 + \varphi_R u_{2R} + \varphi_L u_{2L} + u_3 + \alpha \tilde{\varphi}_0.$$

Then we repeat the process and split u_4 according to

$$u_4 = \varphi_R u_{2R}^{(2)} + \varphi_L u_{2L}^{(2)} + u_3^{(2)} + \alpha^{(2)} \varphi_0 + u_4^{(2)}$$

and define the second iterate by $\hat{u}^{(2)} = u_4 - u_4^{(2)}$. The process is continued, and we have

$$u_4^{(n-1)} = \varphi_R u_{2R}^{(n)} + \varphi_L u_{2L}^{(n)} + u_3^{(n)} + \alpha^{(n)} \varphi_0 + u_4^{(n)}.$$

For l large enough, the size of the forcing in the equation for $u_4^{(n)}$ is reduced in each step and we have

$$\lim_{n \to \infty} \|u_4^{(n)}\|_{L^\infty} = 0.$$

In conclusion, we have constructed a solution of the resolvent equation (4):

$$\hat{u} = \sum_{n=1}^{\infty} \hat{u}^{(n)}$$

for $s \in \Omega_c$. Summing the estimates of the iterates using (22), (26), and (40) we obtain

$$\|\hat{u}\|_{H^1} \le K \left(\|\hat{h}\|_{L^1} + \|\hat{h}\|_{L^2} \right),$$

which completes the proof of Lemma 1.

Appendix. Results for ODEs on a half-line. Here we derive four lemmas on systems of ODEs, with nearly constant coefficients, on a half-infinite interval. These results are used in sections 2.2 and 3. In section A.1 we introduce the ODE and state all assumptions on the coefficients and forcing in the section. In section A.2, on the other hand, we state the results for exactly the problem treated in the main body of the article. Consequently we refer to objects defined in earlier sections.

In this appendix we state the results without proof. The proofs are rather straightforward. Similar results are given in [9], and in [11] these results are given with very minor changes. Proofs can be found in both these references.

A.1. Systems without a parameter. Consider the ODE

$$u_x = Au + f, \qquad x \in [0, \infty), \tag{41}$$

where the coefficient matrix converges to a constant matrix

$$\lim_{x\to\infty} A(x) = A_0.$$

The limiting matrix has no purely imaginary eigenvalues, so we can block-diagonalize it according to

$$TA_0S = \begin{pmatrix} A^+ & 0 \\ 0 & A^- \end{pmatrix},$$

where A^+ is a $k \times k$ matrix. We denote by T^I the first k rows of T and by T^{II} the last $n-k$ rows. We have the following measure of the difference between the coefficient matrix and the limiting matrix:

$$\gamma = \int_0^\infty |A(x) - A_0| dx.$$

We are now in position to state two lemmas.

LEMMA 3. *Consider* (41) *with the boundary conditions*

$$T^{II}u(0) = u_0^{II} \tag{42}$$

and assume $f \in L^1$. Then there exists a γ_0, determined by A_0, such that, if $\gamma \leq \gamma_0$, there exists a unique solution of the problem (41), (42). *This solution satisfies the following estimate:*

$$\|u\|_{L^\infty} + \|u\|_{L^1} + \|u\|_{L^2} \leq K\left(|u_0^{II}| + \|f\|_{L^1}\right),$$

where the constant K depends only on A_0.

LEMMA 4. *Consider* (41) *with $f = 0$. Then there is a γ_0 such that, if $\gamma < \gamma_0$, there is a matrix P such that all bounded solutions u of* (41) *satisfy*

$$(T^I - PT^{II})u(0) = 0.$$

We also have $|P| \leq K\gamma$. The constants γ_0 and K depend only on A_0.

A.2. Systems with a parameter. We consider

$$\begin{pmatrix} u \\ v \end{pmatrix}_x = D(x,s)\begin{pmatrix} u \\ v \end{pmatrix} + f, \qquad x \in [l,\infty), \quad s \in \Omega_c, \tag{43}$$

where

$$D(x,s) = \begin{pmatrix} B^{-1}A & B^{-1} \\ sI & 0 \end{pmatrix},$$

and B, A, and Ω_c denote the same matrices and region, respectively, as in the main body of the article. For this problem we have the following difference between the coefficient matrix and the limiting matrix D_0:

$$\int_l^\infty |D(x,s) - D_0(s)|dx \le Ke^{-\beta l}.$$

This is because the shock profile approaches the left and right states exponentially in x.

We now state two lemmas for the system (43), corresponding to the two lemmas in the preceding section.

Lemma 5. *Consider* (43) *with* $f \in L^1$ *and boundary conditions*

$$\begin{pmatrix} T^{II}(s) \\ T^{IV}(s) \end{pmatrix}\begin{pmatrix} u(l) \\ v(l) \end{pmatrix} = \begin{pmatrix} u_0^{II} \\ v_0^{II} \end{pmatrix}.$$

Then there is an l_0*, independent of* s*, such that for* $l > l_0$*, there is a unique bounded solution. This solution satisfies the estimates*

$$\|u\|_{L^\infty} + \|v\|_{L^\infty} \le K\left(|u_0^{II}| + |v_0^{II}| + \|f\|_{L^1}\right), \tag{44}$$

$$\|u\|_{H^1} + \|v\|_{H^1} \le \frac{K}{|s|}\left(|u_0^{II}| + |v_0^{II}| + \|f\|_{L^1}\right). \tag{45}$$

The constant K *in these inequalities depends only on* B *and* A_R*. The norms are taken over the interval* $[l,\infty)$*.*

Remark. The estimates for the H^1-norm of the solution break down in the limit $s \to 0$ (for $s \in \Omega_c$). This is a consequence of the fact that some of the eigenvalues κ_i of D_0 tend to zero as $s \to 0$ (see (6)), so the decay rate is not uniform in s. When this lemma is applied in the derivation of the resolvent estimate, it is possible to get, an s-independent estimate because we have first reduced the forcing to $\mathcal{O}(s)$.

Lemma 6. *Consider* (43) *with* $f = 0$*. Then there is an* l_0 *such that, for all* $l > l_0$*, all bounded solutions of* (43) *satisfy*

$$\tilde{R}(l,s)\begin{pmatrix} u(l) \\ v(l) \end{pmatrix} = 0, \tag{46}$$

where $\tilde{R}$ *can be written*

$$\tilde{R}(l,s) = \begin{pmatrix} T_R^I(s) \\ T_R^{III}(s) \end{pmatrix} - e^{-\beta l}\tilde{Q}(l,s)\begin{pmatrix} T_R^{II}(s) \\ T_R^{IV}(s) \end{pmatrix}.$$

$\tilde{Q}$ *is analytic in* s *and can be written*

$$\tilde{Q}(l,s) = \begin{pmatrix} P(l) & Q_{12}(l) \\ 0 & 0 \end{pmatrix} + \mathcal{O}(s),$$

where P was introduced in (12) *and Q_{12} is a bounded function of l. Conversely, any function $(u, v)^T$ which satisfies* (43) *and* (46) *is bounded.*

Remark. Using the expression (8) for $T_R(s)$ we obtain

$$\tilde{R}(l, s) = \begin{pmatrix} T^I_{2R} - e^{-\beta l} P(l) T^{II}_{2R} & \tilde{R}_{12}(l) \\ 0 & T^I_{1R} \end{pmatrix} + sQ_1(s, l), \tag{47}$$

where $\tilde{R}_{12}$ is a bounded function of l and Q_1 is a bounded function of s and l, for $l > l_0$ and $s \in \Omega_c$. We observe that the submatrix in the upper left corner is just the matrix $R(l)$ from (11).

REFERENCES

[1] M. Bultelle, M. Grassin, and D. Serre, *Unstable Godunov discrete profiles for steady shock waves*, SIAM J. Numer. Anal., 35 (1998), pp. 2272–2297.

[2] J. J. Erpenbeck, *Stability of step shocks*, Phys. Fluids, 5 (1962), pp. 1181–1187.

[3] H. Freistühler and K. Zumbrun, *Examples of Unstable Viscous Shock Waves*, Institut für Mathematik, RWTH, Aachen, Germany, 1998.

[4] J. Goodman, *Nonlinear asymptotic stability of viscous shock profiles for conservation laws*, Arch. Ration. Mech. Anal., 95 (1986), pp. 325–344.

[5] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[6] K. Knopp, *Theory of Functions* II. *Applications and Continuation of the General Theory*, Dover Publications, New York, 1947.

[7] N. Kopell and L. N. Howard, *Bifurcations and trajectories joining critical points*, Adv. Math., 18 (1975), pp. 306–358.

[8] G. Kreiss, *Convergence to steady state of solutions of viscous conservation laws*, in Proceedings of the Fourth International Conference on Hyperbolic Problems, A. Donato and F. Oliveira, eds., Notes Numer. Fluid Mech. 43, Vieweg, Braunschweig, 1993, pp. 377–384.

[9] G. Kreiss and H.-O. Kreiss, *Stability of systems of conservation laws*, Comm. Pure Appl. Math., 51 (1998), pp. 1397–1424.

[10] P. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 11, SIAM, Philadelphia, 1973.

[11] M. Liefvendahl, *On the Stability of Shock Waves*, Tech. Report TRITA-NA-9908, KTH, Stockholm, Sweden, 1999.

[12] M. Liefvendahl and G. Kreiss, *Examples of Unstable Viscous Shock Waves*, Tech. Report TRITA-NA-0014, KTH, Stockholm, Sweden, 2000.

[13] T. P. Liu, *Nonlinear stability of shock waves for viscous conservation laws*, Mem. Amer. Math. Soc., 56 (1985), pp. 1–108.

[14] A. Majda, *The stability of multidimensional shock fronts*, Mem. Amer. Math. Soc., 41 (1983), pp. 1–95.

[15] A. Majda and R. L. Pego, *Stable viscosity matrices for systems of conservation laws*, J. Differential Equations, 56 (1985), pp. 229–262.

[16] A. Matsumura and K. Nishihara, *On the stability of travelling wave solutions of a one-dimensional model system for compressible viscous gas*, Japan J. Appl. Math., 2 (1985), pp. 17–25.

[17] A. Szepessy and Z. Xin, *Nonlinear stability of viscous shock waves*, Arch. Ration. Mech. Anal., 122 (1993), pp. 53–103.

[18] K. Zumbrun and P. Howard, *Pointwise semigroup methods and stability of viscous shock waves*, Indiana Univ. Math. J., 47 (1998), pp. 741–871.

SIAM J. Math. Anal.
Vol. 33, No. 4, pp. 930–945

CRITICAL THRESHOLDS IN A CONVOLUTION MODEL FOR NONLINEAR CONSERVATION LAWS*

HAILIANG LIU† AND EITAN TADMOR†

Abstract. In this work we consider a convolution model for nonlinear conservation laws. Due to the delicate balance between the nonlinear convection and the nonlocal forcing, this model allows for narrower shock layers than those in the viscous Burgers' equation and yet exhibits the conditional finite time breakdown as in the damped Burgers' equation. We show the critical threshold phenomenon by presenting a lower threshold for the breakdown of the solutions and an upper threshold for the global existence of the smooth solution. The threshold condition depends only on the relative size of the minimum slope of the initial velocity and its maximal variation. We show the exact blow-up rate when the slope of the initial profile is below the lower threshold. We further prove the L^1 stability of the smooth shock profile, provided the slope of the initial profile is above the critical threshold.

Key words. wave breakdown, critical threshold, shock profile, stability

AMS subject classifications. Primary, 35L65; Secondary, 35B30

PII. S0036141001386908

1. Introduction.

Consider the scalar equation of the form

$$u_t + uu_x = Q * u - u, \tag{1.1}$$

where Q is a regular symmetric kernel, monotonically decreasing on $\mathbb{R}^+$, subject to initial data

$$u(0,x) = u_0(x), \quad u_0 \in C_b^1(\mathbb{R}). \tag{1.2}$$

We are concerned with the critical threshold phenomenon supported by the balance between the nonlinear convection and the nonlocal source term in (1.1).

For the kernel Q, we make the following assumption:

(H1) $Q \in C^1(\mathbb{R})$, $Q(-r) = Q(r) \geq 0$, $\int Q(y)dy = 1$, $\int Q(y)|y|dy < \infty$, and $Q'(x) \leq 0$ for $x \geq 0$.

To clarify the effect of the nonlocal term on the right-hand side of (1.1), we make a hyperbolic scaling

$$(t,x) \to \left(\frac{t}{\epsilon}, \frac{x}{\epsilon}\right), \quad \epsilon > 0,$$

which leads to

$$u_t + uu_x = \frac{1}{\epsilon}[Q_\epsilon * u - u], \tag{1.3}$$

where $Q_\epsilon := \frac{1}{\epsilon}Q(\frac{x}{\epsilon})$ and is converging to a delta function $\delta(x)$ as the scaled parameter ϵ tends to zero.

*Received by the editors March 27, 2001; accepted for publication (in revised form) July 17, 2001; published electronically December 28, 2001. This research was supported in part by ONR grant N00014-91-J-1076 and by NSF grant DMS01-07917.

http://www.siam.org/journals/sima/33-4/38690.html

†Department of Mathematics, UCLA, Los Angeles, CA 90095-1555 (hliu@math.ucla.edu, tadmor@math.ucla.edu).

A typical example of the kernel Q is $\frac{1}{2}e^{-|x|}$; with this specific kernel, (1.3) can be written as

$$u_t + uu_x = \mathcal{F}^{-1}\left[\frac{-\epsilon\xi^2}{1+\epsilon^2\xi^2}\hat{u}(t,\xi)\right] = \epsilon\mathcal{F}^{-1}\left[\frac{1}{1+\epsilon^2\xi^2}\hat{u}(t,\xi)\right]_{xx}, \tag{1.4}$$

which is called an R-C-E model after Rosenau's regularized version of the Chapman–Enskog expansion for hydrodynamics [17]. The operator on the right-hand side of (1.4) looks like the usual viscosity term ϵu_{xx} at low wave-number ξ, while for higher wave numbers it is intended to model a bounded approximation of a linearized collision operator, thereby avoiding the artificial instabilities that occur when the Chapman–Enskog expansion is truncated after a finite number of terms [17]. This idea has been greatly advanced recently by Slemrod and his collaborators. A renormalization procedure was introduced in [19] to eliminate the truncation instability and to produce the desired dissipation; the corresponding applications can be found in [20, 21, 22]. The regularization of the Burnett equations via relaxation was investigated by Jin and Slemrod [5, 6]. The rigorous analysis of the model (1.4), including the existence of the shock profiles, the smoothness, as well as the upper-Lipschitz continuity, has been studied by Schochet and Tadmor [23]. We remark that, as observed in [23], the solution sequence $\{u^\epsilon\}$ of (1.4) does not satisfy the Kružkov entropy inequality. The convergence of the solution u^ϵ of (1.4) to the entropy solution of the inviscid Burgers' equation was proved in [23] via the L^1 contraction argument.

(1.3) with $Q = \frac{1}{2}e^{-|x|}$ can also be written as a hyperbolic-elliptic system

$$u_t + uu_x = \phi_x, \quad x \in \mathbb{R}, \quad t > 0, \tag{1.5}$$

$$\epsilon^2\phi_{xx} - \phi + \epsilon u_x = 0. \tag{1.6}$$

It is easy to see that (1.6) enables one to express ϕ in terms of u formally as

$$\phi = (1-\epsilon^2\partial_x^2)^{-1}\epsilon u_x = \epsilon Q_\epsilon * u_x,$$

which in turn gives the right-hand side of (1.3),

$$\phi_x = \epsilon Q_\epsilon * u_{xx} = \frac{1}{\epsilon}[Q_\epsilon * u - u].$$

The system of equations (1.5)–(1.6) is derived as the third-order approximation of the full system describing the motion of radiating gas in therm-nonequilibrium, while the second-order approximation gives the viscous Burgers' equation $u_t + uu_x = \epsilon u_{xx}$, and the first-order approximation gives the inviscid Burgers' equation $u_t + uu_x = 0$. Hamer [4] studied these equations in the physical respect, especially for the steady progressive shock wave solutions. Noting that if ϵ in (1.6) is small, one has $\phi \sim \epsilon u_x$, which leads to the usual viscous Burgers' equation. The viscous Burgers' equation admits smooth shock wave profiles but does not allow the finite time breakdown. On the other hand, if the parameter ϵ is large, one finds from (1.6) that $\epsilon\phi_{xx} + u_x \sim 0$, which when combined with (1.5) gives the damped Burgers' equation $u_t + uu_x = -u/\epsilon$. This damped equation reflects the conditional breakdown in finite time but does not support monotone traveling waves (shock profiles).

The parameter ϵ in (1.3) does not play a role in our analysis and so will be set to 1 for convenience. Equation (1.3) with $\epsilon = 1$, i.e., (1.1), is a physical model that allows for the shock wave profile and yet exhibits the finite time breakdown. For stability of shock profiles via energy method we refer to [11, 8]. The global weak solution to (1.1) was studied in [23].

As is known, the typical well-posedness result asserts that either a solution of a time-dependent PDE exists for all time (global existence of the smooth solution) or else there is a finite time (called life span) such that some norm of the solution becomes unbounded as the life span is approached (called finite time breakdown). The natural question is whether there is a critical threshold for the initial data such that the global existence of the smooth solution or the finite time breakdown depends only on crossing such a critical threshold. This remarkable critical threshold phenomenon was first observed and studied in [3] for a class of Euler–Poisson equations. In this paper we confirm such a critical threshold phenomenon for (1.1)–(1.2) by giving an upper threshold for the global existence of the smooth solution and a lower threshold for the finite time breakdown. We also show the exact blow-up rate as the life span is approached.

In this paper we shall use the following notation for $g \in L^\infty(\mathbb{R})$ to denote the maximal variation:

$$V(g) := \max_{x\in\mathbb{R}} g(x) - \min_{x\in\mathbb{R}} g(x).$$

The first result tells us the critical threshold phenomenon in (1.1).

THEOREM 1.1. *Consider the Cauchy problem* (1.1)–(1.2) *with initial data* $u_0 \in C_b^1(\mathbb{R})$. *Let the kernel* Q *satisfy* (H_1); *then we have the following:*

- *If* $V(u_0) < \frac{1}{4Q(0)}$ *and*

$$\inf_{x\in\mathbb{R}} \partial_x u_0(x) > -\frac{1}{2}\left[1 + \sqrt{1 - 4Q(0)V(u_0)}\right],$$

then the smooth solution exists for all time.

- *If*

$$\inf_{x} \partial_x u_0(x) < -\frac{1}{2}\left[1 + \sqrt{1 + 4Q(0)V(u_0)}\right],$$

then the solution u *must break down at finite time* T. *Moreover,*

$$\lim_{t\to T}(\min_{x\in\mathbb{R}}\{u_x(t,x)\}) = -\infty$$

and the exact blow-up rate is

$$\lim_{t\to T}((T-t)\min_{x\in\mathbb{R}}\{u_x(t,x)\}) = -1.$$

Concerning this theorem, several remarks are in order.

Remarks. 1. The above results show that the solution behavior of (1.1)–(1.2) depends on the relative size of the minimum slope of the initial profile and its maximal variation. If either the maximal variation is too large or the initial velocity slope is too negative, the solution would lose smoothness in finite time. This peculiar phenomenon explains the result obtained in [23], in which additional constraints on the shock strength are imposed to ensure the smoothness of the shock profiles. Further relation between the smoothness of the shock profiles and the shock strength are given in [8]. The critical threshold phenomenon was already partially observed in previous studies; see [23] and [9].

2. As an example, we take $u_0^\theta(x) = \exp(-x^2/\theta)$ for $\theta > 0$. Note that

$$\inf_{x\in\mathbb{R}} [\partial_x u_0^\epsilon(x)] = -\sqrt{\frac{2}{e\theta}}, \quad V(u_0^\theta) = 1.$$

Therefore, choosing θ so small that

$$\theta < \frac{4}{e(1 + 2Q(0) + \sqrt{1 + 4Q(0)})},$$

we see that $\partial_x u_0^\theta$ is below the lower threshold, and thereby the corresponding solution $u^\theta(t, x)$ breaks down in finite time.

3. Note that at the blow-up time, the solution is still bounded, and the gradient of the solution becomes unbounded from below. Such a breakdown is referred to as wave breaking in the context of the shallow water waves. In [25] Whitham emphasized that wave breaking phenomena are some of the most intriguing long-standing problems of water theory. This issue was first settled recently in [15] for Whitham's equation. Another shallow water equation derived recently by Camassa and Holm [2] can be written as (1.5) coupled with the following equation:

$$\phi_{xx} - \phi - u^2 - \frac{1}{2}u_x^2 = 0.$$

This equation as a completely integrable system has a soliton solution and yet exhibits finite time breakdown phenomena for a large class of initial data, which has been observed and justified by Holm [2], Constantin and Escher [1], and McKean [14]. The main tool used in the above papers is to trace the solution gradient along a curve on which the minimum of the gradient is obtained. In this work we trace the dynamics of the solution gradient along the characteristics, which are well known in the context of the hyperbolic equations; see, e.g., [12, 7, 13]. For the global weak solution to the above shallow water equation, we refer to [24] and references therein.

4. From the results above we see that if the magnitude of the initial profile is small, both thresholds given in Theorem 1.1 are close to $\inf_{x\in\mathbb{R}} \partial_x u_0(x) = -1$, which is exactly the critical threshold for the damped Burgers' equation:

$$u_t + uu_x = -u.$$

Indeed, along the particle path $x(\alpha, t)$ defined by

$$\frac{d}{dt}x(\alpha, t) = u(t, x(\alpha, t)), \quad x(\alpha, 0) = \alpha, \quad \alpha \in \mathbb{R},$$

the gradient of the solution to the damped Burgers' equation above can be written explicitly as

$$u_x(t, x) = [e^t(1 + (\partial_x u_0(\alpha))^{-1}) - 1]^{-1},$$

which is bounded from below for all time if and only if

$$\inf_{x\in\mathbb{R}} \partial_x u_0(x) \geq -1.$$

This remarkable critical threshold phenomenon explains why (1.1) admits narrower shock layers than those in the viscous Burgers' equation. We now turn to discussing the asymptotic behavior of solutions, as the initial data are above the critical threshold. We shall concentrate on the case $u_0(-\infty) = u_- > u_+ = u_0(+\infty)$. As shown in [23], (1.1) with $Q = \frac{1}{2}e^{-|x|}$ admits a smooth shock profile $U(x-st)$ connecting u_+ to u_- if and only if the strength $|V(U)| = |u_+ - u_-| \leq \sqrt{2}$. Considering the conservative form of the equation, the natural question is whether this shock profile is stable in $L^1(\mathbb{R})$.

Our stability result is summarized below.

Theorem 1.2. *Let $U(x - st)$ be a continuous shock profile of* (1.1) *and $S(t)u_0$ be a solution to* (1.1)–(1.2) *with initial data $u_0 \in U + L^1(\mathbb{R})$ and $u_0 \in [\inf U, \sup U]$. If $\partial_x u_0 \geq -\frac{1}{2}[1 + \sqrt{1 - 4Q(0)V(u_0)}]$, then there exists a constant k such that*

$$\lim_{t\to\infty} \|S(t)u_0 - U(\cdot - st + k)\|_{L^1} = 0.$$

Remarks. 1. The $L^p (1 \leq p < \infty)$ stability is immediate from the above L^1 stability result and the L^∞ boundedness of $S(t)u_0$. Consult [8] for the stability of traveling waves via the energy principle.

2. We assume that the initial data are above the upper critical threshold to ensure the regularity of the ω-limit set of the solution. This condition is expected to be relaxed since our upper threshold is not sharp.

We now conclude this section by outlining the rest of the paper. In section 2, we recall several properties of (1.1) and give the estimate of the nonlocal term in (1.1), which paves the way for the next sections. The lower threshold for finite time breakdown is given in section 3, in which we also prove the exact blow-up rate. The upper threshold for global existence of the smooth solution is carried out in section 4. The final section is devoted to the L^1 stability of the shock profiles.

2. Preliminaries.

This section is devoted to some estimates which will be used in the next two sections.

In order to formulate the problem, we denote the solution operator of (1.1) as $S(t)$, indexed with $t \in [0, \infty)$,

$$S(t) : L^\infty(\mathbb{R}) \to L^\infty(\mathbb{R}), \quad t \geq 0,$$

such that the solution $u(t, x)$ of (1.1) with initial data a can be expressed as

$$u(t) = S(t)a.$$

We recall from [23] that the solution operator $S(t)$ satisfies the following properties:

- (translate invariance) $S(t)a(x + k) = (S(t)a)(x + k)$ for any $k \in \mathbb{R}$;
- (conservative) if $a - b \in L^1(\mathbb{R})$, then for all $t > 0$, $S(t)a - S(t)b \in L^1(\mathbb{R})$ and $\int(S(t)a - S(t)b) = \int(a - b)$;
- (L^1 contraction) if $a - b \in L^1(\mathbb{R})$, then $S(t)a - S(t)b \in L^1(\mathbb{R})$ and $\|S(t)a - S(t)b\|_1$ is nonincreasing for $t > 0$;
- (monotonicity) if $a(x) \geq b(x)$ for $x \in \mathbb{R}$, then $S(t)a \geq S(t)b$ for all $t > 0$.

The above monotonicity immediately gives us the following maximum principle.

Lemma 2.1. *Let $u_0 \in L^\infty(\mathbb{R})$. Then the solution $u(t, \cdot)$ is also bounded with*

$$\min_{x\in\mathbb{R}} u_0(x) \leq u(t, \cdot) \leq \max_{x\in\mathbb{R}} u_0(x).$$

This maximum principle leads to the following bounds, which will be used in figuring out our threshold conditions.

Lemma 2.2. *Let u be the smooth solution in $[0,T]$. Then it holds that*

$$\min_{x\in\mathbb{R}} u_0(x) \le Q * u(t,\cdot) \le \max_{x\in\mathbb{R}} u_0(x), \quad t\in[0,T], \tag{2.1}$$

$$-Q(0)V(u_0) \le Q * u_x(t,\cdot) \le Q(0)V(u_0). \tag{2.2}$$

Proof. The first inequality follows from the fact $Q*1=1$ and the L^∞ bound $\min_{x\in\mathbb{R}} u_0(x) \le u(t,\cdot) \le \max_{x\in\mathbb{R}} u_0(x)$. We shall prove the second inequality as follows:

$$\begin{aligned} Q * u_x &= \int_{\mathbb{R}} Q(x-y)u_y(t,y)dy \\ &= \int_{\mathbb{R}} Q_x(x-y)u(t,y)dy \\ &= \left[\int_{-\infty}^{x} Q_x(x-y)u(t,y)dy + \int_{x}^{+\infty} Q_x(x-y)u(t,y)dy\right] \\ &\le \min_{x\in\mathbb{R}} u_0(x)\int_{-\infty}^{x} Q_x(x-y)dy + \max_{x\in\mathbb{R}} u_0(x)\int_{x}^{+\infty} Q_x(x-y)dy \\ &\le Q(0)\left[-\min_{x\in\mathbb{R}} u_0(x) + \max_{x\in\mathbb{R}} u_0(x)\right] = Q(0)V(u_0). \end{aligned}$$

The lower bound $-Q(0)V(u_0)$ is clear from the above estimate. □

The existence of T is ensured by the local existence theorem stated in the following lemma.

Lemma 2.3. *Consider the Cauchy problem* (1.1)–(1.2) *with initial data $u_0 \in C_b^1(\mathbb{R})$. Then there exists a positive constant T, depending only on $\|u_0\|_{C_b^1(\mathbb{R})}$, such that* (1.1)–(1.2) *has a unique smooth solution in $C_b^1(\mathbb{R}\times[0,T])$.*

The proof of this local existence is standard via an iteration scheme; the details are omitted. This local existence provides a base for extending the solution or justifying the finite time breakdown.

3. Blow-up criterion—lower threshold.

This section is devoted to a general discussion of wave breaking criteria.

Theorem 3.1. *Consider the Cauchy problem* (1.1)–(1.2). *The maximal existence time T is finite if and only if the gradient of the solution becomes unbounded from below in finite time.*

Proof. From the local existence in Lemma 2.3 it follows that if the gradient of the solution becomes unbounded from below in finite time, then $T<\infty$.

Let the life span $T<\infty$ and assume that for some constant $M>0$ we have

$$u_x(t,x) \ge -M, \quad (t,x)\in[0,T)\times\mathbb{R}. \tag{3.1}$$

On the other hand, by [23, Theorem 5.1] the solution $u(t,x)$ satisfies the one-sided Lipschitz condition, i.e.,

$$u_x(t,x) \le \frac{1}{(\max_{x\in\mathbb{R}} u_{0x})^{-1}+t} \le \max_{x\in\mathbb{R}} u_{0x} < \infty.$$

Therefore the standard continuation argument enables us to extend the solution to $[0,T+\delta)$ with $\delta>0$, and thereby one must have $T=\infty$. This contradiction ensures that

$$\lim_{t\to T-}(\min_{x\in\mathbb{R}} u_x(t,x)) = -\infty. \qquad □$$

The lower threshold is given in the following theorem.

THEOREM 3.2. *Consider the Cauchy problem* (1.1)–(1.2) *with the initial profile* $u_0 \in C_b^1(\mathbb{R})$*. If* u_0 *is bounded and its gradient is negative with*

$$\inf_{x\in\mathbb{R}} \partial_x u_0(x) < -\frac{1}{2}\big[1+\sqrt{1+4Q(0)V(u_0)}\big],$$

then the life span T *must be finite. Moreover,*

$$T \le \left[-\frac{1}{2}\big(1+\sqrt{1+4Q(0)V(u_0)}\big) - \inf_{x\in\mathbb{R}} \partial_x u_0(x)\right]^{-1}$$

and

$$\lim_{t\to T}(\min_{x\in\mathbb{R}}\{u_x(t,x)\}) = -\infty.$$

Proof. Differentiation of (1.1) with respect to x leads to

$$d_t + ud_x + d^2 = Q * u_x - d, \quad t \in (0,T),$$

where $d := u_x(t,x)$. The smoothness of u ensures that there exists a smooth curve $x(\alpha,t)$ satisfying

$$\frac{d}{dt}x(\alpha,t) = u(t,x(\alpha,t)), \quad x(\alpha,0)=\alpha, \quad \alpha \in \mathbb{R}.$$

Evaluating the above $d-$ equation at $x(\alpha,t)$ and using $Q * u_x \le A := Q(0)V(u_0)$ stated in Lemma 2.2, we have

$$d' + d^2 = Q * u_x(t,x(\alpha,t)) - d \le A - d, \quad ' := \partial_t + u\partial_x$$

for $t \in (0,T)$. That is,

$$d' \le -(d-M_1)(d-M_2), \quad t \in (0,T), \tag{3.2}$$

with

$$M_1 := -\frac{1}{2}[1+\sqrt{1+4A}], \quad M_2 := -\frac{1}{2}[1-\sqrt{1+4A}].$$

For a fixed $\alpha \in \mathbb{R}$, if $d_0(\alpha) := u_0'(\alpha) < M_1$, then we claim that

$$d(t) < d_0(\alpha), \quad t \in (0,T). \tag{3.3}$$

If this would not be true, there is some $t_0 \in (0,T)$ with $d(t) < d_0$ on $[0,t_0)$ and $d(t_0) = d_0$ by the continuity of $d = u_x$ in time. But in this case

$$d' \le -(d_0 - M_1)(d_0 - M_2) < 0, \quad t \in (0,t_0).$$

An integration over $(0,t_0)$ yields

$$d(t_0) < d_0,$$

which contradicts our assumption that $d(t_0) = d_0$ for $t_0 < T$. This implies that (3.3) holds.

Combining (3.3) with (3.2), we obtain

$$d' \leq -(d - M_1)^2, \quad t \in (0, T),$$

and integration yields

$$d(t) \leq M_1 + \left[t - \frac{1}{M_1 - d_0}\right]^{-1}.$$

From this we find that $d(t) \to -\infty$ before t reaches $\frac{1}{M_1 - d_0}$. This proves that the solution breaks down in finite time once $\partial_x u_0 \geq M_1$ fails. □

The blow-up rate at the breaking time is summarized in the next theorem.

THEOREM 3.3. *Let T be the maximal existence time of* (1.1)–(1.2). *If the life span T is finite, then*

$$\lim_{t \to T} \left((T - t) \left(\min_{x \in \mathbb{R}} \{u_x(t, x)\} \right) \right) = -1.$$

Proof. By Theorem 3.1 one has

$$\lim_{t \to T} \left(\min_{x \in \mathbb{R}} \{u_x(t, x)\} \right) = -\infty.$$

For $t \in [0, T)$ the solution u is smooth and the curve $x(\alpha, t)$ is well defined by

$$\frac{d}{dt} x(\alpha, t) = u(t, x(\alpha, t)), \quad x(\alpha, 0) = \alpha, \quad \alpha \in \mathbb{R}.$$

This implies

$$\frac{\partial}{\partial \alpha} x(\alpha, t) = \exp\left(\int_0^t u_x(\tau, x(\alpha, \tau)) d\tau \right) > 0, \quad t \in (0, T),$$

and hence $x(\alpha, t)$ is a one-to-one mapping from $\mathbb{R}$ to $\mathbb{R}$. From these facts it follows that there exists an $\alpha \in \mathbb{R}$ such that

$$\min_{x \in \mathbb{R}} \{u_x(t, x)\} = u_x(t, x(\alpha, t)).$$

As done previously, we consider dynamics of $d = u_x$ along the curve $x(\alpha, t)$, using $-A \leq Q * u_x \leq A = Q(0)V(u_0)$ to obtain

$$-A - d \leq d' + d^2 \leq A - d, \quad t \in (0, T).$$

Let $\epsilon \in (0, 1)$ be suitably small. Since $\lim_{t \to T} d(t) = -\infty$, there exists $t_0 \in (0, T)$ such that

$$d(t) < B^-(\epsilon), \quad t \in [t_0, T), \tag{3.4}$$

with

$$B^-(\epsilon) = \frac{-2A}{\sqrt{1 + 4A\epsilon(2 - \epsilon)} - 1}$$

being the smaller root of $(\epsilon^2 - 2\epsilon)d^2 - d + A = 0$. Otherwise there exists $\delta > 0$ such that

$$d(t) < B^-(\epsilon), \quad t \in (t_0, t_0 + \delta),$$

and for $\delta < T - t_0$

$$d(t_0 + \delta) = B^-(\epsilon).$$

Hence for $d(t) < B^-(\epsilon)$ on $(t_0, t_0 + \delta)$,

$$\frac{d}{dt} d(t) \le A - d - d^2 \le -(1-\epsilon)^2 d^2 < 0, \quad t \in (t_0, t_0 + \delta).$$

Integration gives

$$d(t_0 + \delta) < d(t_0) < B^-(\epsilon).$$

This contradiction shows that

$$d \le B^-(\epsilon), \quad t \in [t_0, T);$$

therefore

$$d' \le -(1-\epsilon)^2 d^2, \quad t \in [t_0, T). \tag{3.5}$$

On the other hand, let

$$B^+(\epsilon) = \frac{-2A}{\sqrt{1 + 4A\epsilon(2+\epsilon)} + 1},$$

which is the bigger root of $(\epsilon^2 + 2\epsilon)d^2 - d - A = 0$. We find that $B^-(\epsilon) < B^+(\epsilon)$ and

$$d(t) < B^+(\epsilon), \quad t \in (t_0, T).$$

This gives $(\epsilon^2 + 2\epsilon)d^2 - d - A > 0$, yielding

$$d' \ge -(d^2 + d + A) \ge -(1+\epsilon)^2 d^2, \quad t \in (t_0, T). \tag{3.6}$$

A combination of (3.5) with (3.6) gives

$$-(1+\epsilon)^2 d^2 \le d' \le -(1-\epsilon)^2 d^2, \quad t \in (t_0, T).$$

Note that d is locally Lipschitz on (t_0, T) and so is $1/d$ on (t_0, T). The above inequality leads to

$$(1-\epsilon)^2 \le \left(\frac{1}{d}\right)' \le (1+\epsilon)^2, \quad t \in (t_0, T).$$

For $t \in (t_0, T)$, integrate the above over (t, T) to obtain

$$-(1-\epsilon)^2 (T - t) \le \frac{1}{d(t)} \le -(1+\epsilon)^2 (T - t), \quad t \in (t_0, T).$$

Optimizing the above in terms of ϵ, one then has

$$\lim_{t \to T} (T - t) d(t) = -1.$$

This completes the proof. □

4. Global smoothness—upper threshold. With the breakdown criterion in section 2, we are ready to discuss the upper threshold for the global existence of the smooth solution to (1.1)–(1.2).

THEOREM 4.1. *Consider the Cauchy problem* (1.1)–(1.2) *with the initial profile* $u_0 \in C_b^1(\mathbb{R})$. *If* u_0 *is bounded with the maximal variation* $V(u_0) \leq \frac{1}{4Q(0)}$ *and its gradient is above an upper threshold, i.e.,*

$$\inf_{x\in\mathbb{R}} \partial_x u_0(x) \geq -\frac{1}{2}\left[1+\sqrt{1-4Q(0)V(u_0)}\right],$$

then the smooth solution exists for all time and satisfies

$$\partial_x u(t,x) \geq -\frac{1}{2}\left[1+\sqrt{1-4Q(0)V(u_0)}\right].$$

Proof. To show the global existence of the smooth solution it suffices to establish an a priori lower bound for the gradient of solution u_x. As argued earlier, we evaluate $d := u_x$ along the particle path $x(\alpha, t)$ to obtain

$$d' + d^2 = Q * u_x(t, x(\alpha,t)) - d(t).$$

Noting that the lower bound of Qu_x is $-A = -V(u_0)Q(0)$, we find that

$$d' \geq -A - d - d^2 = -(d-A_1)(d-A_2),$$

where

$$A_1 = -\frac{1}{2}[1+\sqrt{1-4A}], \quad A_2 = -\frac{1}{2}[1-\sqrt{1-4A}].$$

Now let q solve the following problem:

$$\frac{d}{dt}q(t) = -(q-A_1)(q-A_2), \quad q(0) = d_0.$$

Then the comparison of the above differential relations yields

$$d - q \geq (d_0 - q(0)) \exp\left(-\int_0^t (d+q+1)d\tau\right) = 0, \quad t > 0.$$

However, q can be solved explicitly as

$$q(t) = \left[A_1 - A_2\frac{d_1 - A_1}{d_0 - A_2}\exp(A_2 - A_1)t\right]\left[1 - \frac{d_1 - A_1}{d_0 - A_2}\exp(A_2 - A_1)t\right]^{-1}.$$

Therefore for $A_2 > d_0 \geq A_1$ one has $d(t) \geq q(t) \geq A_1$; for $d_0 \geq A_2$ one has $d(t) \geq q(t) \geq A_2$. The possible breakdown occurs only when $d_0 < A_1$ because

$$q(t^*) = -\infty, \qquad t^* = \frac{1}{A_2 - A_1}\log\frac{d_1 - A_2}{d_0 - A_1} > 0.$$

The lower bound of d cannot be ensured for $d_0 < A_1$. However, $d_0 \geq A_1$ is sufficient to ensure the global existence of the smooth solution. □

5. L^1 stability of shock profiles. Let us rewrite (1.1) as

$$u_t + f(u)_x = Q * u - u, \quad f = u^2/2. \tag{5.1}$$

A shock wave with speed $s \in \mathbb{R}$ is a solution of (5.1) of the form $U(x - st)$, with U approaching two different shock states $u_\pm$ at far field. The function U formally satisfies the equation

$$-sU' + f(U)' = Q * U - U, \quad U(\pm\infty) = u_\pm.$$

The critical threshold phenomenon revealed in the previous sections suggests that the smooth shock profile is possibly subject to some constraints on the shock strength.

Indeed the existence of the shock profiles for (5.1) with convex flux function f has been proved [23, Theorem 3.1], which we state below, for $Q = \frac{1}{2}e^{-|x|}$, for the reader's convenience.

THEOREM 5.1. *Assume $f'' > 0$. Then the Lax shock condition*

$$f'(u_+) < s < f'(u_-) \tag{5.2}$$

and the Rankine–Hugoniot shock condition

$$H(u_+) = 0, \quad H(u) \equiv -s(u - u_-) + f(u) - f(u_-), \tag{5.3}$$

are necessary conditions for the existence of a traveling wave solution

$$U(z \equiv x - st), \quad \lim_{z\to\pm\infty} U(z) = u_\pm,$$

for (5.1). *Conversely, if* (5.2) and (5.3) *hold, then a sufficient condition for the existence of such a traveling wave is*

$$4 \sup_{u_+<u<u_-} \{-f''(u)H(u)\} \le 1,$$

and a necessary condition is

$$4\{-f''(u^*)H(u^*)\} \le 1.$$

Here u^* is defined by

$$f'(u^*) = s.$$

Note that for the Burgers' flux $f = u^2/2$, the shock speed by the Rankine–Hugoniot relation (5.3) becomes $s = \frac{u_+ + u_-}{2}$. If the shock condition (5.2), i.e.,

$$u_+ < u_-$$

holds, then there exists such a traveling wave if and only if

$$|u_+ - u_-| \le \sqrt{2}. \tag{5.4}$$

This shows that the traveling wave solutions of the R-C-E equation give narrower shock layers than those of the viscous Burgers' equation.

Recall that the solution operator

$$S(t) : L^\infty(\mathbb{R}) \to L^\infty(\mathbb{R}), \quad t \ge 0,$$

satisfies the nice properties listed in section 2, which ensures that $S(t)$ can be well extended to $L^1(\mathbb{R}) + L^\infty(\mathbb{R})$ and preserves all those properties.

To reformulate the stability problem, we introduce the following set:

$$A := U + L^1(\mathbb{R}),$$

which is a complete metric space with the metric

$$\rho(a_1, a_2) = \|a_1 - a_2\|_1.$$

We also set two subspaces of A,

$$A_1 := \{U(\cdot + k), \quad k \in \mathbb{R}\}$$

and

$$A_2 = \{a \in A : \lim_{t\to\infty} S(t)a \quad \text{exists and} \lim_{t\to\infty} S(t)a \in A_1\}.$$

Equipped with the above notations, we see that proving the stability result in Theorem 1.2 reduces to proving the relation

$$A \cap [u+, u_-] \subset A_2, \tag{5.5}$$

provided $S(t)a$ is smooth.

We introduce the ω-limit set of a as

$$\omega(a) = \cap_{s\geq 0}\overline{\cup_{t\geq s}\{S(t)a\}}.$$

This ω-limit set is invariant for $S(t)$. In fact, the definition implies that $b \in \omega(a)$ if and only if there is a sequence $\{t_k\} \to \infty$ such that

$$\rho(S(t_k)a, b) \to 0.$$

The following lemma plays a critical role in proving (5.5).

LEMMA 5.2. *If $a, b \in A \cap [u_+, u_-]$ and $a - b$ does not keep same sign on $\mathbb{R}$, then*

$$\|S(t)a - S(t)b\|_1 < \|a - b\|_1, \quad t > 0.$$

Proof. By Kružkov's argument [10] we have

$$\int_0^T \int_{\mathbb{R}} \{|u - v|\phi_t + sgn(u - v)[f(u) - f(v)]\phi_x\} dxdt$$
$$\geq \int_0^T \int_{\mathbb{R}} \{|u - v| - sgn(u - v)G * (u - v)\}\phi dxdt,$$

where ϕ is an arbitrary nonnegative test function. Thus, by taking $\phi(x, t) = \chi(t)\psi(x, t)$, letting $\psi = 1 - g_\epsilon(|x - x_0| - M(T - t))$ with $M = sup|f'|$ tend to the function that is identically one, and letting $\chi(t)$ approximate the indicator function of the interval $[0, t]$, we conclude

$$\|a - b\|_1 - \|S(t)a - S(t)b\|_1 \geq \int_{\mathbb{R}} |S(t)a - S(t)b| - sgn(a - b)G * (S(t)a - S(t)b)dx. \tag{5.6}$$

Using the monotonicity of $S(t)$ we see that if $a-b$ changes sign on $\mathbb{R}$, then so does $S(t)a - S(t)b$. Note that since $\|Q\|_1 = 1$, we find that

$$\int_{\mathbb{R}} |u| - sign(u) Q * u dx = 0$$

if and only if u does not change sign or $u \equiv 0$. This shows that the right-hand side of (5.6) is positive if $a-b$ changes sign on $\mathbb{R}$. □

Armed with the above lemma we proceed to complete the stability proof via the following steps, which have become standard since the work by Osher and Ralston [16] and Serre [18].

First we restrict our stability proof to the initial data in

$$N(U,k_1,k_2) := \{a \in A, \quad U(x+k_1) \le a(x) \le U(x+k_2), \quad \text{for some} \quad k_1, k_2 \in \mathbb{R}\},$$

and we can later extend our argument to a larger class using the following dense lemmas.

Step 1 (dense argument). We first show that both A_1 and A_2 are complete subspaces of A.

Lemma 5.3. *Let U be the monotone shock profile; then A_i, $i = 1, 2$, are close in A.*

Proof. We first show the closeness of A_1. It is easy to see that for any $k \in \mathbb{R}$, $U(x+k) \in A$ since

$$\|U(\cdot + k) - U(\cdot)\|_1 = |k(u_+ - u_-)| < \infty.$$

We assume $U(x+k_n)$ converges in A; then it is a Cauchy sequence. Note that

$$\|U(\cdot + k_n) - U(\cdot + k_m)\|_{L^1} = |(k_n - k_m)(u_+ - u_-)|$$

implies k_n is also a Cauchy sequence in $\mathbb{R}$. Let its limit be k; then by letting $m \to \infty$ in the above equation, one finds that the limit of $U(x+k_n)$ is $U(x+k) \in A_1$.

We now turn to showing the closeness of A_2. Let $a_k \in A_2$ be a Cauchy sequence with its limit being $a \in A$. We need to show $a \in A_2$. Note that for each $a_k \in A_2$ we have that $\lim_{t\to\infty} S(t)a_k = \tilde{a}_k \in A_1$ exists. Hence $\tilde{a}_k$ is a Cauchy sequence in the complete metric space A_1, for

$$\|\tilde{a}_k - \tilde{a}_l\|_1 = \lim_{t\to\infty} \|S(t)a_k - S(t)a_l\|_1 \le \|a_k - a_l\|_1.$$

We denote the limit of $\tilde{a}_k$ by $\tilde{a}$ as $k \to \infty$, which, when combined with the closeness of A_1, implies that $\tilde{a} \in A_1$. Therefore $a \in A_2$ since

$$\|S(t)a - \tilde{a}\|_1 \le \|S(t)a - S(t)a_k\|_1 + \|S(t)a_k - \tilde{a}_k\|_1 + \|\tilde{a}_k - \tilde{a}\|_1 \to 0$$

as $k \to \infty$ and $t \to \infty$. □

Lemma 5.4. *For any given $k_1, k_2 \in \mathbb{R}$, the set $N(U, k_1, k_2)$ is dense in $A \cap [u_+, u_-]$.*

The proof can be done as in [16]; the details are omitted.

Step 2 (compact criteria).

Lemma 5.5. *For any $k_1, k_2 \in \mathbb{R}$, the ω-limit set $\omega(N(U, k_1, k_2))$ is not empty.*

Proof. It suffices to show that $\cup_{t\ge0}\{S(t)a\}$ is precompact for any $a \in N(U, k_1, k_2)$. Indeed, due to $a - U \in L^1$ and the L^1 contraction of $S(t)$ we have

$$\|S(t)a - U\|_1 = \|S(t)a - S(t)U\|_1 \le \|a - U\|_1 < \infty, \quad t \ge 0.$$

The L^1 equicontinuity follows from the fact that

$$\|S(t)a(x+h) - S(t)a(x)\|_1 \le \|a(x+h) - a(x)\|_1 \to 0$$

uniformly in time as h goes to zero. Using the semigroup property of $S(t)$, we have

$$U(x+k_1) \le S(t)a \le U(x+k_2), \quad t \ge 0.$$

Hence

$$||S(t)a - U(x)||_{L^1(|x|>M)} \le \max\{\|U(\cdot + k_1) - U\|_{L^1(|x|>M)}, \|U(\cdot + k_2) - U\|_{L^1(|x|>M)}\} \to 0$$

uniformly in t as M goes to ∞.

When recalling the Frechet–Kolmogorov–Riesz compactness theorem, the above facts yield that $\cup_{t\ge 0}\{S(t)a\}$ is precompact. □

Step 3 (time-invariance).

LEMMA 5.6. *Let $b \in \omega(N(U,k_1,k_2))$. Then for any given $k \in \mathbb{R}$*

$$\|b - U(\cdot + k)\|_1 = \|S(t)b - U(\cdot + k)\|_1.$$

Proof. Since $b \in \omega(N(U,k_1,k_2))$, we see that there exists $a \in N(U,k_1,k_2)$ and a sequence $\{t_n\}$ such that $t_n \to \infty$ as $n \to \infty$ and

$$\lim_{n\to\infty} \|S(t_n)a - b\|_1 = 0.$$

Given any $k \in \mathbb{R}$, by contraction of $S(t)$ we know that

$$\|S(t)a - U(x+k)\|_1 = \|S(t)a - S(t)U(x+k)\|_1$$

is decreasing in time and thus admits a limit $c_k \ge 0$ as $t \to \infty$, i.e.,

$$\lim_{t\to\infty} \|S(t)a - U(x+k)\|_1 = c_k \ge 0.$$

Letting $t = t_n$ in the above equation and passing to the limit, we have

$$\|b - U(\cdot + k)\|_1 = c_k.$$

Note that if $b \in \omega(a)$, then $S(t)b \in \omega(a)$ (ω is invariant under the flow); thereby

$$\|S(t)b - U(\cdot + k)\|_1 = c_k.$$

Therefore

$$\|S(t)b - U(\cdot + k)\|_1 = \|b - U(\cdot + k)\|_1 \quad \forall t > 0, \quad k \in \mathbb{R}. \qquad \square$$

We are now ready to prove (5.5). We first prove

$$N(U,k_1,k_2) \subset A_2.$$

By Lemma 5.5 we know that $\omega(N(U,k_1,k_2))$ is not empty. For $a \in N(U,k_1,k_2)$ and $b \in \omega(a)$, we need to show that there exists a $k \in \mathbb{R}$ such that

$$b = U(x+k).$$

Lemma 5.6 shows that

$$\|b - U(\cdot + k)\|_1 = \|S(t)b - U(\cdot + k)\|_1 = c_k.$$

Noting that $U(x+k)$ is the fixed point of $S(t)$, Lemma 5.2 shows that $b - U(x+k)$ must stay with one sign.

Therefore, choosing

$$k = \int_{\mathbb{R}} (a - U)dx/(u_+ - u_-)$$

gives

$$c_k = \int_{\mathbb{R}} [b - U(\cdot + k)] = \int_{\mathbb{R}} [a - U(\cdot + k)] = 0.$$

On the other hand, since the initial data a are assumed to be above the critical threshold, $\partial_x(S(t)a)$ is uniformly bounded with respect to t, and hence b is Lipschitz continuous. This regularity combined with the above fact yields

$$b = U(x+k).$$

We now conclude the proof of (5.5). Let $a \in A \cap [u_+, u_-]$. We need to show $a \in A_2$.

Using Lemma 5.4 shows that there exists $a_n \in N(U, k_1, k_2) \in A$ such that $\|a_n - a\|_1 \to 0$ as $n \to \infty$. By the above proved fact we see that there exists k_n such that

$$\lim_{t\to\infty} \|S(t)a_n - U(\cdot + k_n)\|_1 = 0.$$

This tells us that $a_n \in A_2$. Due to the closeness of A_2, the limit a also belongs to A_2. Therefore there exists a k such that

$$\lim_{t\to\infty} \|S(t)a - U(\cdot + k)\|_1 = 0;$$

as argued above, the constant k as the limit of $\int_{\mathbb{R}}(a_n - U)dx/(u_+ - u_-)$ is

$$\int_{\mathbb{R}} (a - U)dx/(u_+ - u_-)$$

since $|\int (a_n - a)dx| \le \|a_n - a\| \to 0$ as $n \to \infty$. This completes the proof of (5.5) and thereby of Theorem 1.2.

REFERENCES

[1] A. Constantin and J. Escher, *Wave breaking for nonlinear nonlocal shallow water equations*, Acta Math., 181 (1998), pp. 229–243.

[2] R. Camassa and D.D. Holm, *An integrable shallow water equation with peaked solitons*, Phys. Rev. Lett., 71 (1993), pp. 1661–1664.

[3] S. Engelberg, H. Liu, and E. Tadmor, *Critical Threshold in Euler-Poisson Equations*, UCLA CAM report 01-07, UCLA, Los Angeles, CA, 2001; also available online from http://www.math.ucla.edu/applied/cam/index.html.

[4] K. Hamer, *Nonlinear effects on the propagation of sound waves in a radiating gas*, Quart. J. Mech. Appl. Math., 24 (1971), pp. 155–168.

[5] S. Jin and M. Slemrod, *Regularization of the Burnett equations for rapid granular flows via relaxation*, Phys. D, 150 (2001), pp. 207–218.

[6] S. Jin and M. Slemrod, *Regularization of the Burnett equations via relaxation*, J. Statist. Phys., 103 (2001), pp. 1009–1033.

[7] F. John, *Formation of singularities in one-dimensional nonlinear wave propagation*, Comm. Pure Appl. Math., 27 (1974), pp. 337–405.

[8] S. Kawashima and S. Nishibata, *Shock waves for a model system of the radiating gas*, SIAM J. Math. Anal., 30 (1999), pp. 95–117.

[9] S. Kawashima and S. Nishibata, *Cauchy problem for a model system of the radiating gas: Weak solutions with a jump and classical solutions*, Math. Models Methods Appl. Sci., 9 (1999), pp. 69–91.

[10] S.N. Kružkov, *First order quasilinaer equations in several independent variables*, Math. USSR Sb., 10 (1970), pp. 217–243 (in Russian).

[11] S. Kawashima and Y. Tanaka, *Asymptotic Behavior of Solutions to the One-Dimensional Model System for Radiating Gas*, unpublished note.

[12] P. Lax, *Development of singularities in the nonlinear waves for quasilinear hyperbolic partial differential equations*, J. Math. Phys., 5 (1964), pp. 611–613.

[13] T.P. Liu, *Development of singularities in the nonlinear waves for quasilinear hyperbolic partial differential equations*, J. Differential Equations, 33 (1979), pp. 92–111.

[14] H.P. Mckean, *Breakdown of shallow water equations*, Asian J. Math., 2 (1998), pp. 867–874.

[15] P. Naumkin and I. Shishmarev, *Nonlinear Nonlocal Equations in the Theory of Waves*, Transl. Math. Monogr. 133, Amer. Math. Soc., Providence, RI, 1994.

[16] S. Osher and J. Ralston, *L^1-stability of traveling waves with application to convective porous media flow*, Comm. Pure Appl. Math., 35 (1982), pp. 737–749.

[17] P. Rosenau, *Extending hydrodynamics via the regularization of the Chapman-Enskog expansion*, Phys. Rev. A, 40 (1989), pp. 7193–7196.

[18] D. Serre, *Stabilité des ondes de choc de viscosité qui peuvent être caracteristiques*, preprint, 1995.

[19] M. Slemrod, *A renormalization method for the Chapman-Enskog expansion*, Phys. D, 109 (1997), pp. 257–273.

[20] M. Slemrod, *Renormalization of the Chapman-Enskog expansion: Isothermal fluid flow and Rosenau saturation*, J. Statist. Phys., 91 (1998), pp. 285–305.

[21] M. Slemrod, *Constitutive relations for monatomic gases based on a generalized rational approximation to the sum of the Chapman-Enskog expansion*, Arch. Ration. Mech. Anal., 150 (1999), pp. 1–22.

[22] M. Slemrod, *Constitutive relations for Rivlin-Ericksen fluids based on generalized rational approximation*, Arch. Ration. Mech. Anal., 146 (1999), pp. 73–93.

[23] S. Schochet and E. Tadmor, *Regularized Chapman-Enskog expansion for scalar conservation laws*, Arch. Rational Mech. Anal., 119 (1992), pp. 95–107.

[24] Z. Xin and P. Zhang, *On the weak solutions to a shallow water equation*, Comm. Pure Appl. Math., 53 (2000), pp. 1411–1433.

[25] G.B. Whitham, *Linear and Nonlinear Waves*, Pure Appl. Math., Wiley-Interscience, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 33, No. 4, pp. 946–958

SPECTRAL APPROXIMATION ORDERS OF RADIAL BASIS FUNCTION INTERPOLATION ON THE SOBOLEV SPACE[*]

JUNGHO YOON[†]

Abstract. In this study, we are mainly interested in error estimates of interpolation, using smooth radial basis functions such as multiquadrics. The current theories of radial basis function interpolation provide optimal error bounds when the basis function ϕ is smooth and the approximand f is in a certain reproducing kernel Hilbert space $\mathcal{F}_\phi$. However, since the space $\mathcal{F}_\phi$ is very small when the function ϕ is smooth, the major concern of this paper is to prove approximation orders of interpolation to functions in the Sobolev space. For instance, when ϕ is a multiquadric, we will observe the error bound $o(h^k)$ if the function to be approximated is in the Sobolev space of smoothness order k.

Key words. radial basis function, interpolation, Sobolev space, positive definite function, multiquadric, "shifted" surface spline

AMS subject classifications. 41A05, 41A15, 41A25, 41A30, 41A63

PII. S0036141000373811

1. Introduction.

Radial basis function interpolation is a very useful and convenient tool for multivariate scattered data approximation problems. Its strengths are as follows: (i) it facilitates the evaluation of the approximant; (ii) the accuracy of approximation is usually very satisfactory provided the approximand f is reasonably smooth; (iii) there is enough flexibility in the choice of basis functions. A function $\phi : \mathbb{R}^d \to \mathbb{R}$ is radial in the sense that $\phi(x) = \Phi(|x|)$, where $|\cdot|$ is the usual Euclidean norm.

Let Π_m denote the subspace of $C(\mathbb{R}^d)$ consisting of all algebraic polynomials of degree less than m on $\mathbb{R}^d$. Suppose that a continuous function $f : \mathbb{R}^d \to \mathbb{R}$ is known only at a set of discrete points $X := \{x_1, \ldots, x_N\}$ in $\Omega \subset \mathbb{R}^d$. Radial basis function interpolation to f on X starts with choosing a basis function ϕ, and then it defines an interpolant by

$$a_{f,X}(x) := \sum_{i=1}^{\ell} \beta_i p_i(x) + \sum_{j=1}^{N} \alpha_j \phi(x - x_j), \tag{1.1}$$

where $p_1, \ldots, p_\ell$ is a basis for Π_m and the coefficients α_j $(j = 1, \ldots, N)$ and β_i $(i = 1, \ldots, \ell)$ are chosen to satisfy the linear system

$$\begin{aligned} a_{f,X}(x_j) &= f(x_j), \quad && j = 1, \ldots, N, \\ \sum_{j=1}^{N} \alpha_j p_i(x_j) &= 0, && i = 1, \ldots, \ell. \end{aligned} \tag{1.2}$$

Here, the set of scattered points X has the nondegeneracy property for Π_m; that is, if $p \in \Pi_m$ and $p(x_j) = 0$, $j = 1, \ldots, N$, then $p = 0$. It guarantees that the interpolation

[*] Received by the editors June 13, 2000; accepted for publication (in revised form) June 12, 2001; published electronically December 28, 2001. This work was supported by grants DoD-N00014-97-1-0806 and DEPSCoR-DAAG55-98-1-0002.

http://www.siam.org/journals/sima/33-4/37381.html

[†] Department of Mathematics, Arizona State University, Tempe, AZ 85287-1804 (yoon@math.la.asu.edu).

method reproduces the polynomial space Π_m, i.e., $a_{p,X} = p$ for any $p \in \Pi_m$. For a wide choice of functions ϕ and polynomial orders m, the existence and uniqueness of the solution of the linear system (1.2) is ensured when ϕ is a conditionally positive definite function (see [M]).

DEFINITION 1.1. *Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a continuous function. We say that ϕ is conditionally positive definite of order $m \in \mathbb{N} := \{1, 2, \ldots\}$ if for every finite set of pairwise distinct points $X = \{x_1, \ldots, x_N\} \subset \mathbb{R}^d$ and for every $\alpha = (\alpha_1, \ldots, \alpha_N) \in \mathbb{R}^N \setminus 0$ satisfying*

$$\sum_{j=1}^{N} \alpha_j p(x_j) = 0, \quad p \in \Pi_m,$$

the quadric form

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \phi(x_i - x_j)$$

is positive definite.

In what follows, we assume $\phi = \Phi(|\cdot|)$ to be conditionally positive definite of order m. Also, the function ϕ is considered as a tempered distribution in $\mathcal{D}'(\mathbb{R}^d)$, and we assume that its Fourier transform $\hat{\phi}$ coincides on $\mathbb{R}^d \setminus 0$ with some continuous function while having a certain type of singularity (necessarily of a finite order) at the origin, i.e., $\hat{\phi}$ is of the form

$$|\cdot|^n \hat{\phi} = F > 0, \quad n \geq 0, \quad \text{and } F \in L_\infty(\mathbb{R}^d).$$

Among many radial basis functions, our major concern is with smooth functions ϕ such as multiquadrics $\phi(x) := c_{m,d}(|x|^2 + \lambda^2)^{m-d/2}$, d odd, $m > d/2$, where $c_{m,d}$ is a suitable constant.

For a given basis function ϕ, there arises a function space

$$\mathcal{F}_\phi := \left\{ f : |f|_\phi := \int_{\mathbb{R}^d} \frac{|\hat{f}(\theta)|^2}{\hat{\phi}(\theta)} d\theta < \infty \right\}, \tag{1.3}$$

which is called reproducing kernel Hilbert space (or "native" space) for ϕ ([MN2] and [WS]). For all $x \in \Omega$, $f \in \mathcal{F}_\phi$, bounds for the interpolation error are usually of the form

$$|f(x) - a_{f,X}(x)| \leq P_{\phi,X}(x)|f|_\phi. \tag{1.4}$$

Here $P_{\phi,X}$ is the *power function* that evaluates the norm of the error functional at x:

$$P_{\phi,X}(x) = \sup_{|f|_\phi \neq 0} \frac{|f(x) - a_{f,X}(x)|}{|f|_\phi}.$$

In fact, when the basis function ϕ is smooth, the interpolation method provides optimal asymptotic decay of errors, but the space $\mathcal{F}_\phi$ is very small. The approximands f need to be extremely smooth for effective error estimates. However, practically, most multivariate scattered data are not arising from extremely smooth functions. An error analysis for the case that the underlying function is reasonably smooth needs to be provided. Thus, the main objective of this paper is to prove asymptotic error bounds of interpolation (by using smooth basis function ϕ) to functions in a larger space, especially in the Sobolev space.

Asymptotic approximation properties are usually quantified by the notion of approximation order. In order to make this notion feasible, we measure the "density" of X (in Ω) by

$$h := h(X;\Omega) := \sup_{x\in\Omega} \min_{x_j\in X} |x - x_j|. \tag{1.5}$$

Here we assume that $\Omega \subset \mathbb{R}^d$ is an open bounded domain with both cone property and Lipschitz boundary. In particular, for a given set X, we adopt the scaled basis functions $\phi_\omega := \phi(\cdot/\omega)$, where

$$\omega := \omega(h)$$

is a parameter depending on h such that $h/\omega \to 0$ as $h \to 0$, and we use the notation

$$s_{f,X}(x) := \sum_{i=1}^{\ell} \beta_i p_i(x) + \sum_{j=1}^{N} \alpha_j \phi_\omega(x - x_j) \tag{1.6}$$

to differentiate from the notation $a_{f,X}$ in (1.1). Then our goal is to provide error estimates of $f - s_{f,X}$ of the following form: Let ϕ be a smooth basis function (e.g., multiquadric). Under some suitable conditions of the parameter ω (e.g., $\omega = h^r$ with $r \in [0,1)$), we will show the asymptotic property

$$\|f - s_{f,X}\|_{L_\infty(\Omega)} = o(h^k), \quad h \to 0,$$

provided that $f \in W_\infty^k(\Omega)$, the L_∞-Sobolev space of smoothness order k. To the writer's knowledge, this is the first paper dedicated to the study of spectral approximation order of interpolation to the functions in the Sobolev space $W_\infty^k(\Omega)$. Indeed, Buhmann and Dyn also explored the spectral convergence order of multiquadric interpolation in the paper [BuD]. However, this result considers interpolants on $h \cdot \mathbb{Z}^d$ under some conditions of the underlying function f, while we work with a finite subset X in Ω.

The reader who is interested in knowing more about the state of the art in the area of radial basis function methods may find it useful to consult with the surveys [Bu], [D], and [P]. Other important sources are the works of Wu and Schaback [WS] and especially those of Madych and Nelson [MN1], [MN2], who developed a theory of interpolation based on reproducing kernel Hilbert spaces. Interpolation by compactly supported basis functions has been studied by Wendland [W].

The following notations will be used throughout this paper. For any $k \in \mathbb{N}$, the Sobolev space is defined by

$$W_p^k(\Omega) := \left\{ f : \|f\|_{k,L_p(\Omega)} := \left(\sum_{|\alpha|_1 \le k} \|D^\alpha f\|_{L_p(\Omega)}^p \right)^{1/p} < \infty \right\}$$

with $1 \le p \le \infty$. Several different function norms are used. When $\mathbf{g}$ is a matrix or a vector, $\|\mathbf{g}\|_p$ indicates its p-norm with $1 \le p \le \infty$. For $x \in \mathbb{R}^d$, $|x| := (x_1^2 + \cdots + x_d^2)^{1/2}$ stands as its Euclidean norm. The Fourier transform of $f \in L_1(\mathbb{R}^d)$ is defined as

$$\hat{f}(\theta) := \int_{\mathbb{R}^d} f(t) \exp(-i\theta \cdot t)\, dt.$$

Also, for a function $f \in L_1(\mathbb{R}^d)$ we use the notation $f^\vee$ for the inverse Fourier transform. In particular, the Fourier transform can be uniquely extended to the space of tempered distributions on $\mathbb{R}^d$.

2. The extension of a function f in $W^k_\infty(\Omega)$. Our analysis in this paper requires the construction of a suitable extension of a given function $f \in W^k_\infty(\Omega)$ to a function on $\mathbb{R}^d$. Indeed, the lengthy assumptions on Ω in section 1 assure the existence of a function on $\mathbb{R}^d$ whose restriction to Ω agrees with f. The following result is cited from literature.

THEOREM 2.1 (Brenner and Scott [BrS]). *Suppose that Ω has a Lipschitz boundary. Then for every function $f \in W^k_p(\Omega)$, there is an extension mapping $E : W^k_p(\Omega) \to W^k_p(\mathbb{R}^d)$ defined for all nonnegative integer k and real numbers p in the range $1 \le p \le \infty$ satisfying $Ef|_\Omega = f$ for all $f \in W^k_p(\Omega)$ and*

$$\|Ef\|_{k,L_p(\mathbb{R}^d)} \le c\|f\|_{k,L_p(\Omega)},$$

where the constant c is independent of f.

The construction of our suitable extension of $f \in W^k_\infty(\Omega)$ to a function on $\mathbb{R}^d$ can be done in two steps. First, according to Theorem 2.1, there exists a function $Ef \in W^k_\infty(\mathbb{R}^d)$ such that $Ef|_\Omega = f$. Second, we let σ_Ω be a C^∞-cutoff function such that $\sigma_\Omega(x) = 1$ for $x \in \Omega$ and $\sigma_\Omega(x) = 0$ for $|x| > r$ with a sufficiently large $r > 0$. Then we define an extension f^o by

$$f^o := \sigma_\Omega Ef.$$

Of course, f^o is compactly supported and $f^o(x) = f(x)$ for $x \in \Omega$. Indeed, for a large part of this paper, we wish to work with f^o and not f. For convenience, we will henceforth write f for f^o. Therefore, here and in what follows, *without great loss*, we assume that an approximand $f \in W^k_\infty(\Omega)$ is supported in a sufficiently large compact set in $\mathbb{R}^d$ such that $f \in W^k_\infty(\mathbb{R}^d)$.

3. Error bounds. In this section, we will provide a (modified) method of error analysis of interpolation to functions in the Sobolev space $W^k_\infty(\Omega)$. In addition, we obtain a sufficient condition for the optimal convergence order $\|f - s_{f,X}\|_{L_\infty(\Omega)} = o(h^k)$ with $f \in W^k_\infty(\Omega)$. For this purpose, we start with finding a mollified function (say, f_H) of a given (underlying) function f. The function f_H is supposed to be in the space $\mathcal{F}_{\phi_\omega}$ in (1.3) and should be a good approximation to f in some sense. In order to define a mollification f_H of f, we use a nonnegative C^∞-cutoff function

$$\sigma : \mathbb{R}^d \to [0,1]. \tag{3.1}$$

Here, for convenience, we assume that the function σ is radially symmetric and supp σ lies in the Euclidean ball $B_1 = \{x \in \mathbb{R}^d : |x| \le 1\}$, and we assume that $\sigma = 1$ on $B_{1/2}$ and $\|\sigma\|_{L_\infty(\mathbb{R}^d)} = 1$. Then, letting $\sigma_\delta := \sigma(\cdot/\delta)$ with $\delta > 0$, we construct two functions f_H and f_T by

$$\begin{aligned} f_H &:= \sigma_\delta(h\cdot)^\vee * f, \\ f_T &:= f - \sigma_\delta(h\cdot)^\vee * f. \end{aligned} \tag{3.2}$$

It clearly follows that

$$f = f_H + f_T.$$

Also, due to the fact that the interpolation operator $s_{f,X}$ is linear, it is useful to split the error $f - s_{f,X}$ as follows:

$$f - s_{f,X} = (f_H - s_{f_H,X}) + (f_T - s_{f_T,X}).$$

Accordingly, this section falls naturally into two parts. In the first, since $f_H \in \mathcal{F}_{\phi_\omega}$, we estimate the term $f_H - s_{f_H,X}$ by applying the well-known method in (1.4). The second part of the section deals with $f_T - s_{f_T,X}$. Our main tool for this case is to use stability results on the interpolation process. Afterward, the final result is stated in Theorem 3.6

From the papers (see, e.g., [WS], [MN2]), we cite the following lemma.

LEMMA 3.1. *Let $a_{X,f}$ in (1.1) be an interpolant to f on $X = \{x_1, \ldots, x_N\}$. Given ϕ and m, for all functions f in the space $\mathcal{F}_\phi$, there is an error bound of the form*

$$|f(x) - a_{f,X}(x)| \le |f|_\phi P_{\phi,X}(x),$$

where $P_{\phi,X}(x)$ is the norm of the error functional at x, i.e.,

$$P_{\phi,X}(x) = \sup_{|f|_\phi \neq 0} \frac{|f(x) - a_{f,X}(x)|}{|f|_\phi}. \tag{3.3}$$

The following lemma estimates the error $f_H - s_{f_H,X}$.

LEMMA 3.2. *Let $f_H := \sigma_\delta(h\cdot)^\vee * f$ with $\sigma_\delta(h\cdot)$ as the cutoff function in (3.1), and let $s_{f_H,X}$ in (1.6) be the interpolant to f_H on X using the basis function ϕ_ω. Let ω be a parameter depending on h, i.e., $\omega = \omega(h)$. Then, for every $f \in L_2(\mathbb{R}^d)$, we have an estimate of the form*

$$|f_H(x) - s_{f_H,X}(x)| \le P_{\phi,X/\omega}(x/\omega) M_{\phi,\omega}(\delta/h) \|f\|_{L_2(\mathbb{R}^d)}, \quad x \in \Omega,$$

where $M_{\phi,\omega}(r)$, $r > 0$, is defined by

$$M_{\phi,\omega}(r) := \sup_{\theta \in B_r} |\hat{\phi}_\omega(\theta)|^{-1/2}. \tag{3.4}$$

Proof. Recalling the definition of $s_{f_H,X}$ in (1.6), one simply notes that the function $s_{f_H,X}(\omega\cdot)$ can be considered as an interpolant (employing the shifts of ϕ) to the scaled function $f_H(\omega\cdot)$ on X/ω, i.e.,

$$s_{f_H,X}(\omega\cdot) = \sum_{i=1}^{\ell} \beta_i p_i(x) + \sum_{j=1}^{N} \alpha_j \phi(\cdot - x_j/\omega) = a_{f_H(\omega\cdot),X/\omega},$$

with $a_{f,X}$ in (1.1). Then, since $f_H(\omega\cdot)$ belongs to the space $\mathcal{F}_\phi$, Lemma 3.1 can be used directly to derive the bound

$$\begin{aligned} |f_H(x) - s_{f_H,X}(x)| &= |f_H(\omega\cdot) - a_{f_H(\omega\cdot),X/\omega}|(x/\omega) \\ &\le P_{\phi,X/\omega}(x/\omega)|f_H(\omega\cdot)|_\phi. \end{aligned} \tag{3.5}$$

Now, in order to estimate the term $|f_H(\omega\cdot)|_\phi$, we find from the definition of f_H in (3.2) that

$$\widehat{f_H(\omega\cdot)}(\theta) = \omega^{-d} \sigma_\delta(h\theta/\omega) \hat{f}(\theta/\omega).$$

Then the explicit formula of the norm $|\cdot|_\phi$ in (1.3) induces by change of variables that

$$|f_H(\omega\cdot)|_\phi^2 = \omega^{-d}\int_{\mathbb{R}^d} \left|\sigma_\delta(h\theta)\hat{f}(\theta)\right|^2 \hat{\phi}^{-1}(\omega\theta)d\theta$$
$$\leq \sup_{\theta\in B_{\delta/h}} |\hat{\phi}_\omega(\theta)|^{-1}\|f\|_{L_2(\mathbb{R}^d)}^2.$$

Due to the expression (3.5), we finish the proof. □

Now, we are going to turn to the estimate of the error $f_T - s_{f_T,X}$ with f_T in (3.2). Since there is no guarantee that the function f_T belongs to the space $\mathcal{F}_\phi$, the classical method of the error analysis of interpolation is not applicable to this case. Hence, in order to make the estimate $f_T - s_{f_T,X}$ feasible, we employ the stability results on interpolation process. To this end, we define the separation distance within X by

$$q := q_X := \min_{1\leq i\neq j\leq N} |x_i - x_j|/2. \tag{3.6}$$

It is well known from literature (e.g., [NSW2], [S1]) that as q is getting smaller, the condition number of the interpolation matrix becomes larger. Also, the irregularity of a set X can be measured by the ratio h/q. In particular, we assume that the sets of scattered points considered in this study are sets of quasi-uniformly distributed points. These sets satisfy the following property: There exists a constant $\eta > 0$ independent of X such that

$$2q \leq h \leq \eta q. \tag{3.7}$$

This condition implies that the number of the scattered points in the set X is bounded above by a quantity that depends on the density of X, i.e., $N = O(h^{-d})$. On the other hand, we particularly introduce a function φ defined by

$$\varphi := \sigma_\epsilon^\vee = \sigma(\cdot/\epsilon)^\vee, \tag{3.8}$$

where σ_ϵ is the cutoff function in (3.1). For the purpose of simplifying the following analysis, we assume $\epsilon > 0$ to be any fixed number satisfying the condition

$$\epsilon < \delta/\eta \tag{3.9}$$

with δ in (3.2). It is obvious that the Fourier transform of φ is $\hat{\varphi} = \sigma_\epsilon$, which is supported in the ball B_ϵ. Furthermore, since σ is a C^∞-cutoff function, $\varphi(x)$ decays fast as x tends to ∞. Indeed, the function φ is employed to use the stability results on the interpolation process. It first requires us to show that φ is a conditionally positive definite radial function. For this proof, we find the following identity:

$$\sum_{i=1}^N\sum_{j=1}^N \alpha_i\alpha_j\varphi(x_i - x_j) = \int_{\mathbb{R}^d} \hat{\varphi}(\theta)\left|\sum_{j=1}^N \alpha_j e^{ix_j\cdot\theta}\right|^2 d\theta$$

for any $\alpha = (\alpha_1,\ldots,\alpha_N) \in \mathbb{R}^N \setminus 0$. Since the map $\theta \mapsto \sum_{j=1}^N \alpha_j e^{ix_j\cdot\theta}$, $\theta \in \mathbb{R}^d$, has zeros at most on a set of measure zero, we see that the integral in the right-hand side of the above identity is always positive. It is asserted from Definition 1.1 that the function φ is conditionally positive definite of order $m = 0$. Also, since the cutoff function σ_ϵ is radially symmetric, its inverse Fourier transform φ is also a radial function (see [S3]). Then an interpolant to f on X using the (scaled) function

$$\varphi_q(x) := \varphi(x/q)$$

is of the form

$$g_{f,X}(x) = \sum_{j=1}^{N} \beta_j \varphi_q(x - x_j). \tag{3.10}$$

One simply notes that the matrix $\mathbf{A}_{\varphi_q} := (\varphi_q(x_i - x_j))_{i,j=1,\ldots,N}$ is positive definite.

Proposition 3.3. *Let X be a q-separated set with q in equation* (3.6). *Let $\mathbf{b}_f := (\beta_1, \ldots, \beta_N)^T$, and let $\mathbf{A}_{\varphi_q} := (\varphi_q(x_i - x_j))_{i,j=1,\ldots,N}$ be the interpolation matrix by φ_q. Then we have the following properties:*

(a) $\|\mathbf{A}_{\varphi_q}^{-1}\|_2 \le c_1$ *for some* $c_1 > 0$.

(b) $\|\mathbf{A}_{\varphi_q}^{-1}\|_1 = \|\mathbf{A}_{\varphi_q}^{-1}\|_\infty \le c_2 \|\mathbf{A}_{\varphi_q}^{-1}\|_2$ *for some* $c_2 > 0$.

(c) $\|\mathbf{b}_f\|_\infty \le c_3 \|f\|_{L_\infty(\mathbb{R}^d)}$ *for some* $c_3 > 0$.

Proof. Since the interpolation matrix $\mathbf{A}_{\varphi_q}$ has the separation distance 1, the matrix norm $\|\mathbf{A}_{\varphi_q}^{-1}\|_2$ is bounded by a constant (see [NSW2]). Furthermore, the basis function φ_q decays fast around ∞, and the inequality in (b) is proved by a direct application of Theorem 3.11 in the paper [BSW]. The identity $\|\mathbf{A}_{\varphi_q}^{-1}\|_1 = \|\mathbf{A}_{\varphi_q}^{-1}\|_\infty$ is an obvious consequence of symmetry. Finally, to prove (c), we find that the matrix $\mathbf{b}_f$ can be written as

$$\mathbf{b}_f = \mathbf{A}_{\varphi_q}^{-1}\mathbf{f}$$

with $\mathbf{f} := (f(x_1), \ldots, f(x_N))^T$. After some direct calculations, one can prove the inequality $\|\mathbf{b}_f\|_\infty \le \|\mathbf{A}_{\varphi_q}^{-1}\|_1 \|f\|_{L_\infty(\mathbb{R}^d)}$. Hence, by using (b), the relation in (c) is immediate. □

Before estimating the error $f_T - s_{f_T,X}$, we cite the following result.

Lemma 3.4 (Yoon [Y1]). *Let $f_T = f - \sigma_\delta(h\cdot)^\vee * f$ with $\sigma_\delta(h\cdot)$ the cutoff function in* (3.1). *Then, for every $f \in W_\infty^k(\mathbb{R}^d)$ with k a positive integer, we have the following decaying property:*

$$\|f_T\|_{L_\infty(\mathbb{R}^d)} = \|f - f_H\|_{L_\infty(\mathbb{R}^d)} = o(h^k).$$

Lemma 3.5. *Let X be a set of scattered points with the condition* (3.7), *and let $s_{f_T,X}$ in* (1.6) *be the interpolant to f_T on X using ϕ_ω, where $f_T = f - \sigma_\delta(h\cdot)^\vee * f$ and $\omega = \omega(h)$. Then, for every $f \in W_\infty^k(\Omega)$ with k a positive integer, there is an error bound of the form*

$$|f_T(x) - s_{f_T,X}(x)| \le o(h^k)(1 + P_{\phi,X/\omega}(x/\omega) M_{\phi,\omega}(\delta/h)), \quad x \in \Omega,$$

as $h \to 0$, with $M_{\phi,\omega}(r)$, $r > 0$, in (3.4).

Proof. Let us first define a function $\tilde{f}$ by

$$\tilde{f} := h^{-k} f_T.$$

It is clear that $h^{-k} s_{f_T,X} = s_{\tilde{f},X}$. Then, we employ the interpolant $g_{\tilde{f},X}$ in (3.10) to derive the following bound:

$$h^{-k}|f_T(x) - s_{f_T,X}(x)| \le |\tilde{f}(x)| + |g_{\tilde{f},X}(x)| + |g_{\tilde{f},X}(x) - s_{\tilde{f},X}(x)|.$$

The convergence property $\|f_T\|_{L_\infty(\mathbb{R}^d)} = o(h^k)$ in Lemma 3.4 yields that $\|\tilde{f}\|_{L_\infty(\mathbb{R}^d)} = o(1)$ as h tends to 0. Also, by applying Proposition 3.3, we get

$$\begin{aligned} |g_{\tilde{f},X}(x)| &\le \|\mathbf{b}_{\tilde{f}}\|_\infty \sum_{j=1}^N \varphi_q(x - x_j) \\ &\le c\,\|\tilde{f}\|_{L_\infty(\mathbb{R}^d)} = o(1). \end{aligned}$$

Here, since X is a q-separated set and the function φ_q decays fast around ∞, we can easily check that $\sum_{j=1}^N \varphi_q(\cdot - x_j)$ is uniformly bounded on Ω. Therefore, it remains to show that the term $g_{\tilde{f},X} - s_{\tilde{f},X}$ is bounded by $o(1)P_{\phi,X/\omega}(x/\omega)M_{\phi,\omega}(\delta/h)$ as $h \to 0$. For this, we claim that

$$s_{\tilde{f},X} = s_{g_{\tilde{f},X},X}.$$

In fact, this identity is immediate from the interpolation property $\tilde{f}(x_j) = g_{\tilde{f},X}(x_j)$ for any $j = 1, \dots, N$. Then, applying the same technique as in the proof of Lemma 3.2 gives us the bound

$$(3.11) \qquad |s_{\tilde{f},X}(x) - g_{\tilde{f},X}(x)| \le P_{\phi,X/\omega}(x/\omega)|g_{\tilde{f},X}(\omega\cdot)|_\phi, \quad x \in \Omega.$$

Moreover, according to the definition of the norm $|\cdot|_\phi$, we get

$$(3.12) \qquad \begin{aligned} |g_{\tilde{f},X}(\omega\cdot)|_\phi^2 &= \int_{\mathbb{R}^d} \left| \sum_{j=1}^N \beta_j e^{ix_j\cdot\theta} \right|^2 \sigma_\epsilon^2(q\theta)\hat{\phi}_\omega^{-1}(\theta) q^{2d} d\theta \\ &\le M_{\phi,\omega}^2(\epsilon/q) \int_{\mathbb{R}^d} \left| \sum_{j=1}^N \beta_j \varphi_q(x - x_j) \right|^2 dx. \end{aligned}$$

Remembering the relations $\frac{1}{q} \le \frac{\eta}{h}$ in (3.7) and $\epsilon < \frac{\delta}{\eta}$ in (3.9), we easily find that $\frac{\epsilon}{q} \le \frac{\eta\epsilon}{h} \le \frac{\delta}{h}$. Then since $M_{\phi,\omega}(r)$ is monotonically increasing as r grows, it follows that

$$(3.13) \qquad M_{\phi,\omega}(\epsilon/q) \le M_{\phi,\omega}(\delta/h).$$

Also, since $\sum_{j=1}^N \varphi_q(\cdot - x_j)$ is uniformly bounded, we have

$$(3.14) \qquad \begin{aligned} \int_{\mathbb{R}^d} \left| \sum_{j=1}^N \beta_j \varphi_q(x - x_j) \right|^2 dx &\le c\|\mathbf{b}_{\tilde{f}}\|_\infty^2 \int_{\mathbb{R}^d} \left| \sum_{j=1}^N \varphi_q(x - x_j) \right| dx \\ &\le c'\|\tilde{f}\|_{L_\infty(\mathbb{R}^d)} = o(1) \end{aligned}$$

by Proposition 3.3 and the condition $N = O(h^{-d})$. Hence, inserting (3.13) and (3.14) into (3.12), we arrive at the bound

$$|g_{f,X}|_\phi \le M_{\phi,\omega}(\delta/h)o(1).$$

Together with (3.11), we complete the proof of this lemma. □

From Lemma 3.2 and Lemma 3.5, we realize that one of the important ingredients for the estimate $f - s_{f,X}$ is the term $M_{\phi,\omega}(\delta/h)$, $\delta > 0$. Observing the definition of

f_H in (3.2) carefully, we find that the number δ can be chosen arbitrarily. Of course, a certain choice of δ should induce a suitable bound of $P_{\phi,X/\omega}(x/\omega)M_{\phi,\omega}(\delta/h)$, which leads to a desirable estimate of $f - s_{f,X}$. We are now ready to describe the main result of this section.

THEOREM 3.6. *Let X be a set of scattered points with the condition* (3.7), *and let $s_{f,X}$ in* (1.6) *be an interpolant to f on X using the basis function $\phi_\omega = \phi(\cdot/\omega)$. Let $M_{\phi,\omega}(r)$, $r > 0$, be defined as in* (3.4). *Assume that there exists a constant $\delta_0 > 0$ such that*

$$P_{\phi,X/\omega}(x/\omega)M_{\phi,\omega}(\delta_0/h) \le o(h^k).$$

Then, for every function $f \in W_\infty^k(\Omega)$ with k a positive integer, we have an error bound of the form

$$\|f - s_{f,X}\|_{L_\infty(\Omega)} = o(h^k).$$

4. Applications to special radial basis functions. We now turn to applications to special radial basis functions. Employing some known basis functions ϕ, we will show that the interpolant $s_{f,X}$ provides optimal approximation orders for $f \in W_\infty^k(\Omega)$. All the examples here are based on Theorem 3.6.

Example 4.1. Let the radial basis function ϕ be chosen to be one of the following:

(a) $\phi_\lambda(x) := (-1)^{\lceil m-d/2\rceil}(|x|^2+\lambda^2)^{m-d/2}$, d odd, $m > d/2$ (multiquadrics),

(b) $\phi_\lambda(x) := (-1)^{m-d/2+1}(|x|^2+\lambda^2)^{m-d/2}\log(|x|^2+\lambda^2)^{1/2}$, $m > d/2$, d even ("shifted" surface splines).

(c) $\phi_\lambda(x) := (|x|^2+\lambda^2)^{m-d/2}$, $0 < m < d/2$ (inverse multiquadrics),

where $d,\ m \in \mathbb{N}$ and $\lambda > 0$, and where $\lceil s\rceil$ indicates the smallest integer greater than s. Note that we stress the parameter λ by using the notation ϕ_λ. We find (see [GS]) that the Fourier transform of ϕ_λ is of the form

$$\hat{\phi}_\lambda = c(m,d)\tilde{K}_m(\lambda\cdot)|\cdot|^{-2m},$$

where $c(m,d)$ is a positive constant depending on m and d, and $\tilde{K}_\nu(|t|) := |t|^\nu K_\nu(|t|) \neq 0$, $t \ge 0$, with $K_\nu(|t|)$ the modified Bessel function of order ν. It is well known from literature (e.g., [AS]) that

$$\tilde{K}_\nu \sim (1 + |\cdot|^{(2\nu-1)/2})\exp(-|\cdot|).$$

Then, for all $\theta \in B_{\delta/h}$, we have the bound $\hat{\phi}_\lambda(\omega\theta)^{-1/2} \le c|\omega\delta/h|^m \exp(\lambda\omega\delta/2h)$ for a constant $c > 0$. It leads to the inequality

$$M_{\phi_\lambda,\omega}(\delta/h) \le c(\delta)\omega^{-d/2}|\omega/h|^m \exp(\lambda\omega\delta/2h),$$

where $c(\delta)$ is a constant depending on δ. On the other hand, due to Madych and Nelson [MN3], we see that there exists a constant $c' > 0$ independent of X such that

$$P_{\phi_\lambda,X/\omega}(x/\omega) \le c\exp(-c'\lambda\omega/h)$$

for a sufficiently small $h > 0$. Since $\omega^{m-d/2} \le o(h^{-d/2})$ (see section 1), from the above two inequalities, we arrive at the expression

$$M_{\phi_\lambda,\omega}(\delta/h)P_{\phi_\lambda,X/\omega}(x/\omega) \le c(\delta)h^{-m-d/2}\exp\left(-\frac{\lambda\omega}{h}(c'-\delta/2)\right). \tag{4.1}$$

Here, we can choose a sufficiently small $\delta_0 > 0$ such that $c' - \delta/2 > 0$ for any $\delta \leq \delta_0$. In particular, we assume ω to satisfy the relation

$$h|\log h|^{1+r} \leq \omega$$

for any fixed $r > 0$. Then, it follows that

$$\exp\left(-\frac{\lambda\omega}{h}\left(c' - \delta_0/2\right)\right) \leq \exp\big(-\lambda|\log h|^{1+r}(c' - \delta_0/2)\big) \\ = h^{\lambda(c'-\delta_0/2)|\log h|^r}. \tag{4.2}$$

Indeed, as h tends to 0, the number $\lambda(c' - \delta_0/2)|\log h|^r > 0$ becomes arbitrarily large. Hence, for any given $k \in \mathbb{N}$, there exists a sufficiently small $h_0 > 0$ such that $h^{\lambda(c'-\delta/2)|\log h|^r} \leq o(h^{k+m+d/2})$ for any $h \leq h_0$. Consequently, together with (4.1) and (4.2), we conclude that

$$M_{\phi_\lambda,\omega}(\delta_0/h)P_{\phi_\lambda,X/\omega}(x/\omega) \leq o(h^k), \quad h \leq h_0.$$

According to Theorem 3.6, we have the following result.

THEOREM 4.1. *Let ϕ_λ be one of the radial basis functions: multiquadrics, inverse multiquadrics, and "shifted" surface splines. Let X be a set of scattered points with the condition* (3.7)*, and let $s_{f,X}$ in* (1.6) *be an interpolant to f on X using $\phi_\lambda(\cdot/\omega)$. Assume that $\omega = \omega(h)$ is chosen to satisfy the relation*

$$h|\log h|^{1+r} \leq \omega$$

for any fixed $r > 0$. Then, for every $f \in W^k_\infty(\Omega)$ with k a positive integer, we have an error estimate of the form

$$\|f - s_{f,X}\|_{L_\infty(\Omega)} = o(h^k) \quad \text{as} \quad h \to 0.$$

COROLLARY 4.2. *Let ϕ_λ be one of the radial basis functions: multiquadrics, inverse multiquadrics, and "shifted" surface splines. Let X be a set of scattered points with the condition* (3.7)*, and let $s_{f,X}$ in* (1.6) *be an interpolant to f on X using $\phi_\lambda(\cdot/\omega)$. Assume that $\omega(h) = h^s$ with $s \in [0, 1)$ or $\omega(h) = h|\log h|^{1+r}$ with $r > 0$. Then, for every $f \in W^k_\infty(\Omega)$ with k a positive integer,*

$$\|f - s_{f,X}\|_{L_\infty(\Omega)} = o(h^k) \quad \text{as} \quad h \to 0.$$

Remark. Recalling that the interpolant $a_{f,X}$ in (1.1) uses the original (nonscaled) basis function, we make an observation concerning the interpolants $a_{f,X}$ in relation to $s_{f,X}$. Given a set X, assume that the interpolant $a_{f,X}$ employs the basis function $\phi_{\omega\lambda}$ instead of ϕ_λ. Then, one should realize that the interpolant $a_{f,X}$ is identically equal to $s_{f,X}$, which uses ϕ_λ. The equality can be verified by the uniqueness of the solution of the linear system (1.2). The reader is referred to the paper [Y2] for the details of the proof.

Example 4.2. Let us consider the basis function ϕ whose Fourier transform $\hat{\phi}$ is of the form

$$\hat{\phi}(\theta) = \exp(-|\theta|^a)$$

with $0 < a \leq 1$. In the case $a = 1$, the basis function ϕ becomes the so-called Poisson kernel

$$\phi = \frac{c_d}{(1 + |\cdot|^2)^{(d+1)/2}}$$

with a suitable constant c_d. For any $\theta \in B_{\delta/h}$, we get $\hat{\phi}(\omega\theta)^{-1} \leq \exp((\omega\delta/h)^a)$. It leads to the inequality

$$M_{\phi,\omega}(\delta/h) \leq \omega^{-d/2} \exp((\omega\delta)^a/h^a).$$

Also, due to Madych and Nelson (see [MN3]), there exists a constant $c' > 0$ independent of X such that

$$P_{\phi,X/\omega}(x/\omega) \leq c \exp(-c'\omega^a/h^a)$$

for sufficiently small $h > 0$. Invoking the condition $\omega^{-d/2} \leq o(h^{-d/2})$, we derive from the above inequalities that

$$M_{\phi,\omega}(\delta/h)P_{\phi,X/\omega}(x/\omega) \leq ch^{-d/2} \exp\left(-\frac{\omega^a}{h^a}(c' - \delta^a)\right). \tag{4.3}$$

Now, in a similar fashion to the case of Example 4.1, we can choose a sufficiently small $\delta_0 > 0$ such that $c' - \delta^a > 0$ for any $\delta \leq \delta_0$. In particular, we assume ω to satisfy

$$h^a |\log h|^{1+r} \leq \omega^a$$

for any fixed $r > 0$. Then, it follows that

$$\begin{aligned}\exp\left(-\frac{\omega^a}{h^a}(c' - \delta_0^a)\right) &\leq \exp\big(-|\log h|^{1+r}(c' - \delta_0^a)\big)\\ &= h^{|\log h|^r(c'-\delta_0^a)}.\end{aligned}$$

Here, $|\log h|^r(c' - \delta_0^a)$ becomes arbitrarily large as h tends to 0. Thus, for any given $k \in \mathbb{N}$, there exists a sufficiently small $h_0 > 0$ such that $h^{|\log h|^r(c'-\delta^a)} \leq o(h^{k+d/2})$ if $h \leq h_0$. Therefore, together with (4.3), we conclude that

$$M_{\phi_\lambda,\omega}(\delta/h)P_{\phi_\lambda,X/\omega}(x/\omega) \leq o(h^k), \quad h \leq h_0.$$

According to Theorem 3.6, we have the following result.

THEOREM 4.3. *Let ϕ be the basis function whose Fourier transform $\hat{\phi}$ is defined by $\hat{\phi} = \exp(-|x|^a)$ with $a \leq 1$. Let X be a set of scattered points with the condition (3.7), and let $s_{f,X}$ in (1.6) be an interpolant to f on X using $\phi_\lambda(\cdot/\omega)$. Assume that $\omega = \omega(h)$ is chosen to satisfy the relation*

$$h|\log h|^{(1+r)/a} \leq \omega$$

with a fixed number $r > 0$. Then, for every $f \in W_\infty^k(\Omega)$ with k a positive integer, we have an error bound of the form

$$\|f - s_{f,X}\|_{L_\infty(\Omega)} = o(h^k) \quad \textit{as} \quad h \to 0.$$

COROLLARY 4.4. *Let ϕ be the function whose Fourier transform $\hat{\phi}$ is of the form $\hat{\phi} = \exp(-|x|^a)$ with $0 < a \leq 1$. Let X be a set of scattered points with the condition (3.7), and let $s_{f,X}$ in (1.6) be an interpolant to f on X using $\phi_\lambda(\cdot/\omega)$. Assume that $\omega = h^s$ with $s \in [0, 1)$ or $\omega = h|\log h|^{(1+r)/a}$ with $r > 0$. Then, for every $f \in W_\infty^k(\Omega)$ with k a positive integer,*

$$\|f - s_{f,X}\|_{L_\infty(\Omega)} = o(h^k) \quad \textit{as} \quad h \to 0.$$

Remark. The Gaussian function $\phi := \exp(-\alpha|\cdot|^2)$, $\alpha > 0$, is not included in the examples of section 4. Indeed, the "quadratic exponential" error bound $\exp(-c/h^2)$, $c > 0$, of its power function $P_{\phi,X}$ is necessary to obtain the condition

$$P_{\phi,X/\omega}(x/\omega)M_{\phi,\omega}(\delta_0/h) \leq o(h^k)$$

for some $\delta_0 > 0$. However, it is not yet proven in the bounded domain case, but it is shown only on all of $\mathbb{R}^d$ under certain circumstances. The reader is referred to the manuscript [S3] for the details. More generally, for any given basis function ϕ, there would be a general theorem on the bounds of $P_{\phi,X}$ in terms of $\hat{\phi}$. In fact, we can easily check that $P_{\phi,X}$ is dependent only on the Fourier transform $\hat{\phi}$, more precisely, on the decaying property of $\hat{\phi}$ (see [WS] and [MN2] for the details).

REFERENCES

[AS] M. Abramowitz and I. Stegun, *A Handbook of Mathematical Functions*, Dover Publications, New York, 1970.

[BSW] B. J. C. Baxter, N. Sivakumar, and J. D. Ward, *Regarding the p-norms of radial basis interpolation matrices*, Constr. Approx., 10 (1994), pp. 451–468.

[BrS] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[Bu] M. D. Buhmann, *New developments in the theory of radial basis functions interpolation*, in Multivariate Approximation: From CAGD to Wavelets, K. Jetter and F. I. Utreras, eds., World Scientific, Singapore, 1993, pp. 35–75.

[BuD] M. D. Buhmann and N. Dyn, *Spectral convergence of multiquadric interpolation*, Proc. Edinburgh Math. Soc. (2), 36 (1993), pp. 319–333.

[D] N. Dyn, *Interpolation and approximation by radial and related functions*, in Approximation Theory VI, Vol. 1, C. K. Chui, L. L. Schumaker, and J. Ward, eds., Academic Press, Boston, 1989, pp. 211–234.

[GS] I. M. Gelfand and G. E. Shilov, *Generalized Functions*, Vol. 1, Academic Press, New York, London, 1964.

[LW] W. Light and H. Wayne, *Error Estimates for Approximation by Radial Basis Functions*, in Wavelet Analysis and Approximation Theory, S. P. Singh and A. Carbone, eds., Kluwer Academic, Dordrecht, 1995, pp. 215–246.

[M] C. A. Micchelli, *Interpolation of scattered data: Distance matrices and conditionally positive functions*, Constr. Approx., 2 (1986), pp. 11–22.

[MN1] W. R. Madych and S. A. Nelson, *Multivariate interpolation and conditionally positive function* I, Approx. Theory Appl., 4 (1988), pp. 77–89.

[MN2] W. R. Madych and S. A. Nelson, *Multivariate interpolation and conditionally positive function* II, Math. Comp., 54 (1990), pp. 211–230.

[MN3] W. R. Madych and S. A. Nelson, *Bounds on multivariate polynomials and exponential error estimates for multiquadric interpolation*, J. Approx. Theory, 70 (1992), pp. 94–114.

[NSW1] F. J. Narcowich, N. Sivakumar, and J. D. Ward, *Norms of inverses and condition numbers for matrices associated with scattered data*, J. Approx. Theory, 64 (1991), pp. 69–94.

[NSW2] F. J. Narcowich, N. Sivakumar, and J. D. Ward, *On condition numbers associated with radial-function interpolation*, J. Math. Anal. Appl., 186 (1994), pp. 457–485.

[P] M. J. D. Powell, *The theory of radial basis functions approximation in* 1990, in Advances in Numerical Analysis, Vol. II: Wavelets, Subdivision Algorithms and Radial Basis Functions, W. A. Light, ed., Oxford University Press, New York, 1992, pp. 105–210.

[S1] R. Schaback, *Error estimates and condition numbers for radial basis function interpolation*, Adv. Comput. Math., 3 (1995), pp. 251–264.

[S2] R. Schaback, *Approximation by radial functions with finitely many centers*, Constr. Approx., 12 (1996), pp. 331–340.

[S3] R. Schaback, *Reconstruction of Multivariate Functions from Scattered Data*, manuscript.

[W] H. Wendland, *Error estimates for interpolation by compactly supported radial basis functions of minimal degree*, J. Approx. Theory, 93 (1998), pp. 258–272.

[WS] Z. Wu and R. Schaback, *Local error estimates for radial basis function interpolation of scattered data*, IMA J. Numer. Anal., 13 (1993), pp. 13–27.

[WW] J. H. Wells and R. L. Williams, *Embeddings and Extensions in Analysis*, Ergeb. Math. Grenzgeb., 84, Springer-Verlag, Berlin, 1975.

[Y1] J. Yoon, *Approximation in $L_p(\mathbb{R}^d)$ from a space spanned by the scattered shifts of radial basis function*, Constr. Approx., 17 (2001), pp. 227–247.

[Y2] J. Yoon, *Interpolation by radial basis functions on Sobolev space*, J. Approx. Theory, 112 (2001), pp. 1–15.

SIAM J. Math. Anal.
Vol. 33, No. 4, pp. 959–981

STABILITY OF L^∞ SOLUTIONS FOR HYPERBOLIC SYSTEMS WITH COINCIDING SHOCKS AND RAREFACTIONS*

STEFANO BIANCHINI†

Abstract. We consider a hyperbolic system of conservation laws

$$\begin{cases} u_t + f(u)_x = 0, \\ u(0, \cdot) = u_0, \end{cases}$$

where each characteristic field is either linearly degenerate or genuinely nonlinear. Under the assumption of coinciding shock and rarefaction curves and the existence of a set of Riemann coordinates w, we prove that there exists a semigroup of solutions $u(t) = \mathcal{S}_t u_0$, defined on initial data $u_0 \in L^\infty$. The semigroup $\mathcal{S}$ is continuous w.r.t. time and the initial data u_0 in the L^1_{loc} topology. Moreover, $\mathcal{S}$ is unique and its trajectories are obtained as limits of wave front tracking approximations.

Key words. hyperbolic systems, conservation laws, well posedness

AMS subject classification. 35L65

PII. S0036141000377900

1. Introduction.

Consider the Cauchy problem for a strictly hyperbolic system of conservation laws

$$\begin{cases} u_t + f(u)_x = 0, \\ u(0, \cdot) = u_0, \end{cases} \tag{1.1}$$

where $u \in \mathbb{R}^n$ and $f : \Omega \mapsto \mathbb{R}^n$ is sufficiently smooth, Ω open. If the initial data u_0 are of small total variation, the global existence was proved first in [18]. Moreover, a series of papers [6, 7, 9, 10, 15] establishes the uniqueness and well posedness of the Cauchy problem (1.1). However, when u_0 has large total variation or, even more generically, u_0 belongs to L^∞, the solution u may not exist globally in L^∞ [20]: only for special systems it is possible to consider initial data with large total variation. We recall some of the results available in this direction.

(1) For scalar conservation laws, the entropy solution to (1.1) generates a contracting semigroup w.r.t the L^1 distance on a domain of L^∞ data [21].

(2) For a general Temple class system, in [3, 5, 24] the existence and stability of the entropy solution for initial data with arbitrarily large but bounded total variation are proved.

(3) If all characteristic families are genuinely nonlinear and the system is Temple class, the existence and stability for initial data in L^∞ are proved in [12].

(4) For special 2×2 systems, in which one of the equations is autonomous, various results have been proved in [4, 16], with initial data with unbounded total variation.

An open question is whether the semigroup of solutions to the systems of case (2), defined on all the initial data u_0 with total variation arbitrary large but bounded, can be extended to data in L^∞. In many systems, in fact, some of the characteristic fields are linearly degenerate, so that the results of [12] do not apply.

*Received by the editors September 8, 2000; accepted for publication (in revised form) July 16, 2001; published electronically December 28, 2001. This research was partially supported by the European TMR Network on Hyperbolic Conservation Laws ERBFMRXCT960033.

http://www.siam.org/journals/sima/33-4/37790.html

†S.I.S.S.A.-I.S.A.S., via Beirut 2-4, 34014 Trieste, Italy (bianchin@sissa.it, http://www.sissa.it/bianchin).

An example is the 2×2 traffic model considered in [2],

$$\left\{ \begin{array}{rcl} \rho_t + \big(\rho v\big)_x & = & 0, \\ \big(\rho(v+p(v))\big)_t + \big(\rho v(v+p(\rho))\big)_x & = & 0, \end{array} \right. \tag{1.2}$$

where $\rho(t,x)$ is the density of cars in the point (t,x) and $v(t,x)$ is their velocity. In this model, the first eigenvalue is genuinely nonlinear and the integral curves of the corresponding right eigenvector are straight lines. The second eigenvector is linearly degenerate, so that the assumption of coinciding shock and rarefaction curves is verified for this system. The existence of a set of Riemann coordinates follows by the fact that the system is 2×2.

Another example is a simple 2×2 model for chromatography,

$$\left\{ \begin{array}{rcl} u^1_t + \left(\dfrac{u^1}{1+u^1+u^2} \right)_x & = & 0, \\ u^2_t + \left(\dfrac{u^2}{1+u^1+u^2} \right)_x & = & 0, \end{array} \right.$$

where all characteristic fields are linearly degenerate and the integral curves of the eigenvalues are straight lines. The major difficulty here in applying the results of [12] is the fact that the total variation of the solution does not decay in time.

The aim of this paper is to prove that, at least in the case where the eigenvalues are genuinely nonlinear or linearly degenerate and shocks and rarefactions coincide, the solution to (1.1) can be defined for $u_0 \in L^\infty$.

This result is particularly interesting from the point of view of control theory. Consider for example the traffic model (1.2) in the quarter plane $t \geq 0$, $x \geq 0$: this system describes the flow of cars in a highway, given a boundary condition $\tilde{u}(t)$ on the line $x = 0$. The function $\tilde{u}$ can be thought of as a control on the system: we are allowed to choose $\tilde{u}$ in order to minimize some prescribed cost functional, for example, the average time spent by a car to arrive from $x = 0$ to $x = \bar{x}$. As shown in [1], in general the compactness of the attainable set can be obtained only with L^∞ boundary data.

To illustrate the heart of the matter, we assume that the system (1.1) admits a system of Riemann coordinates $w \in \mathbb{R}^n$, and that shock and rarefaction curves coincide in Ω. Moreover, we assume that each characteristic field is linearly degenerate or genuinely nonlinear. Differently from [11], we do not assume that rarefaction curves are straight lines. We consider a set E of the form

$$E \doteq \Big\{ u \in \Omega : \ w(u) \in [a_i, b_i], \ i = 1, \dots, n \Big\}.$$

With $L^\infty(\mathbb{R}; E)$ we denote the space of L^∞ functions with values in E. The main result of this paper is the following.

THEOREM 1.1. *There exists a unique semigroup $\mathcal{S} : [0, +\infty) \times L^\infty(\mathbb{R}; E) \longmapsto L^\infty(\mathbb{R}; E)$ such that the following properties are satisfied:*

(i) *for all $u_n, u \in L^\infty(\mathbb{R}; E)$, $t_n, t \in [0, +\infty)$, with $u_n \to u$ in L^1_{loc}, $|t - t_n| \to 0$ as $n \to +\infty$,*

$$\lim_{n \to +\infty} \mathcal{S}_{t_n} u_n = \mathcal{S}_t u \quad \textit{in } L^1_{\mathrm{loc}};$$

(ii) *the trajectory $\mathcal{S}_t u_0$ is a weak entropy solution to the Cauchy problem* (1.1) *for every $u_0 \in L^\infty(\mathbb{R}; E)$;*

(iii) *if u_0 is piecewise constant, then, for t sufficiently small, $\mathcal{S}_t u_0$ coincides with the function obtained by piecing together the solutions of the corresponding Riemann problems.*

From the results of [11, 14], any solution to (1.1) satisfying Lax entropy conditions and a weak regularity assumption is unique. Theorem 1.1 proves that it is possible to define a weak solution $u(t)$ when the initial data are in L^∞ so that $u(t)$ depends continuously w.r.t. the initial data u_0. The uniqueness follows because $\mathcal{S}$ satisfies (iii) and it is the limit of wave front approximations.

As is shown in the last example of [12], the semigroup $\mathcal{S}$ cannot be uniformly continuous; thus we cannot apply any compactness argument to construct the solution $u(t) \doteq \mathcal{S}_t u_0$. The fundamental problem is that, differently from [12], the total variation of the Riemann invariants corresponding to linearly degenerate families does not decrease in time.

The main idea of this paper is to study how the solution to the characteristic equation

$$\dot{x}(t) = \lambda_i(u(t,x(t))), \quad x(0) = y, \tag{1.3}$$

depends on the solution u of (1.1). Denote with $x(t,y)$ the solution of (1.3).

It will be shown that, for a fixed time τ, the map $y \mapsto x(\tau,y)$ depends Lipschitz continuously on the initial data u_0, and, moreover, the Lipschitz constant is independent of the total variation of u_0. Since the Riemann invariant w_i is the broad solution to

$$(w_i)_t + \lambda_i(u(t,x))(w_i)_x = 0, \tag{1.4}$$

a simple argument gives the convergence of the wave front tracking approximations. We recall that a broad solution of (1.4) with initial data $\bar{w}_i(\cdot)$ is given by $w_i(x(t,y)) = \bar{w}_i(y)$, where $x(t,x)$ is the solution to (1.3). In other words, the value of w_i is constant along the integral lines of (1.3).

We note that the stability of the map $y \mapsto x(t,y)$ implies also the well posedness of the ODE (1.3) when $u(t,x)$ is an L^∞ solution of the system (1.1). This result is quite surprising because, as noted in [16], for general hyperbolic systems the solution to (1.3) does not exist or it is not unique. In our case, the assumption on the existence of Riemann invariants and the conservation form of (1.1) implies the continuous dependence of $x(t,y)$ on the initial data u_0, and then we can extend the notion of solutions to (1.3) when u_0 is in L^∞.

The paper is organized as follows. Section 2 contains the basic assumptions on system (1.1). Moreover, we construct the wave front approximation of the solution $u(t)$. In section 3 we carefully analyze the shift differential map, i.e., the evolution of a perturbation in u_0 in which only the position of the initial jumps has changed. The method we use is essentially the one in [12], with slight modifications due to the fact that in our system the rarefaction curves do not need to be straight lines. The main result here is the explicit computation of the shift differential map.

Section 4 is concerned with the equation for characteristics (1.3). We prove the Lipschitz dependence of the map $y \mapsto x(t)$ w.r.t. both the initial data u_0 and y. Moreover, we will show that the Lipschitz constant is independent from the total variation of u_0. Finally, in section 5, we prove Theorem 1.1.

2. Basic assumptions and wave front approximations.

We consider a strictly hyperbolic system of conservation laws

$$u_t + f(u)_x = 0, \tag{2.1}$$

where $f : \Omega \to \mathbb{R}^n$ is a smooth vector field defined on some open set $\Omega \subseteq \mathbb{R}^n$. Let $A(u) \doteq Df(u)$ be the Jacobian matrix of f and denote with $\lambda_i(u)$ its eigenvalues and with $r_i(u)$, $l^i(u)$ its right and left eigenvectors, respectively. We assume that the eigenvalues λ_i can be either genuinely nonlinear or linearly degenerate. In the following the ith rarefaction curve through $u \in \Omega$ will be written as $R_i(s)u$, with $R_i(0)u = u$, while the ith shock curve will be denoted by $S_i(s)u$, and its speed by $\sigma_i(s,u)$. The directional derivative of a function $\phi(u)$ in the direction of $r_i(u)$ will be denoted as

$$r_i \bullet \phi(u) \doteq \lim_{h\to 0} \frac{\phi(u + hr_i(u)) - \phi(u)}{h},$$

while the left and right limit of a bounded variation function f in a point x will be written as

$$f(x-) = \lim_{y\to x-} f(y), \qquad f(x+) = \lim_{y\to x+} f(y).$$

We assume that the rarefaction curves R_i generate a system of Riemann coordinates $w(u)$. We recall that a necessary and sufficient condition for the local existence of Riemann coordinates is the Frobenius involutive condition: if $[X,Y]$ denotes the Lie bracket of the vector fields X,Y, the condition is

$$[r_i, r_j] \in \operatorname{span}\{r_i, r_j\} \quad \text{for all } i, j = 1, \dots, n.$$

In the following we will use indifferently the conserved coordinates u or the Riemann coordinates w.

Fix a domain

$$E \doteq \Big\{ u \in \Omega : \ w(u) \in [a_i, b_i], \ i = 1, \dots, n \Big\}. \tag{2.2}$$

Since E is compact, there is a constant $c > 0$ such that

$$r_i \bullet \lambda_i(u) > c \quad \text{for all } u \in E \quad \text{if } \lambda_i \text{ is genuinely nonlinear.} \tag{2.3}$$

We suppose that the system (2.1) is uniformly strictly hyperbolic in Ω: this means that there exists a constant d such that

$$\lambda_{i+1}(u) - \lambda_i(v) \geq d \quad \text{for all } u, v \in E, \ i = 1, \dots, n-1. \tag{2.4}$$

We also assume that in the system (2.1) shock and rarefaction curves coincide: this implies [27] that either the rarefaction curve $R_i(s)u$ is a straight line or the eigenvalue is linearly degenerate. In fact, one can prove that

$$\frac{d^2}{ds^2}\sigma_i(s,u)\bigg|_{s=0} = \frac{1}{6}(r_i \bullet \lambda_i(u))\langle l^i(u), r_i \bullet r_i(u)\rangle + \frac{1}{3} r_i \bullet (r_i \bullet \lambda_i(u)), \tag{2.5}$$

and for the shock curve $S_i(s)u$ we have

$$\langle l^j(u), S'''(0)u - R'''(0)u\rangle = \frac{1}{2(\lambda_j(u) - \lambda_i(u))}(r_i \bullet \lambda_i(u))\langle l^j(u), r_i \bullet r_i(u)\rangle. \tag{2.6}$$

If λ_i is genuinely nonlinear, the left-hand side of (2.6) is zero if and only if the rarefaction curve is a straight line, because $r_i \bullet r_i(u)$ is orthogonal to $r_i(u)$.

The flux function f thus satisfies the following assumptions:

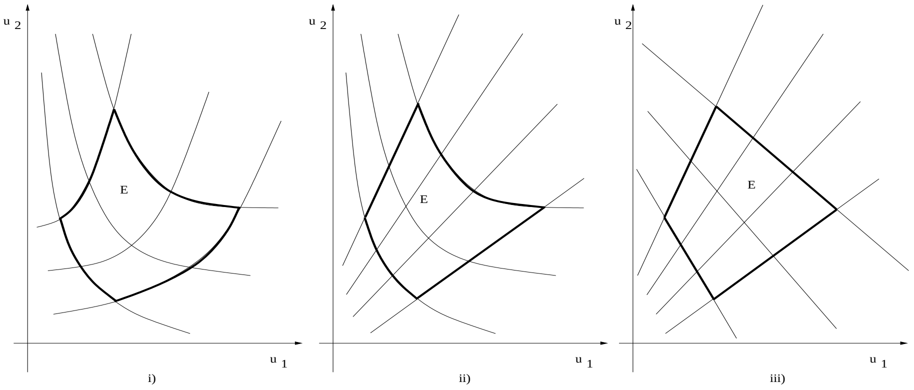


FIG. 1. *The various situations for a* 2×2 *system considered in Remark* 2.1.

(H1) the eigenvalues λ_i of Df are linearly degenerate or genuinely nonlinear;

(H2) the rarefaction curves form a system of coordinates;

(H3) shock and rarefaction curves coincide.

The system (2.1) thus has n_{ld} linearly degenerate fields λ_i, corresponding to the Riemann invariants w_i, and $n_{gnl} = n - n_{ld}$ genuinely nonlinear fields λ_k, corresponding to the Riemann invariants w_k. In the latter case we have $r_k \bullet r_k(u) = 0$ for all $u \in E$.

REMARK 2.1. *If* $\Omega \subseteq \mathbb{R}^2$*, then the rarefaction curves* $R_i(s)u$ *always generate a system of Riemann coordinates. Thus our assumptions are satisfied by the following classes of systems:*

(i) *both eigenvalues are linearly degenerate;*

(ii) *one eigenvalue is linearly degenerate, the other genuinely nonlinear, and the rarefaction curves of the latter are straight lines;*

(iii) *both eigenvalues are genuinely nonlinear and the system is of Temple class.*

The various situations are shown in Figure 1. *Case* (ii) *corresponds to the traffic model considered in* [2]*, while case* (i) *corresponds to* 2×2 *chromatography.*

Given the two points $u^-, u^+ \in E$, with coordinates $u^- = u(w_1^-, \dots, w_n^-)$ and $u^+ = u(w_1^+, \dots, w_n^+)$, with $w_i^+ \neq w_i^-$, consider the intermediate states $u(\omega_i)$, where

$$\omega_0 = w(u^-), \quad \omega_i = (w_1^+, \dots, w_i^+, w_{i+1}^-, \dots, w_n^-), \quad i = 1, \dots, n. \tag{2.7}$$

For all $i = 1, \dots, n$, we denote with $v_i(u^-, u^+)$ the vectors defined as

$$v_i(u^-, u^+) = u(\omega_i) - u(\omega_{i-1}), \tag{2.8}$$

and we define $r_i(u^-, u^+)$ as

$$r_i(u^-, u^+) = \begin{cases} \dfrac{v_i(u^-, u^+)}{|v_i(u^-, u^+)|} = \dfrac{u(\omega_i) - u(\omega_{i-1})}{|u(\omega_i) - u(\omega_{i-1})|} & \text{if } w_i^- \neq w_i^+, \\ r_i(\omega_{i-1}) = r_i(\omega_i) & \text{if } w_i^- = w_i^+, \end{cases} \tag{2.9}$$

where $r_i(u)$ is the ith eigenvector of $DF(u)$. We assume that the vectors $r_i(u^-, u^+)$ are linearly independent for all $u^-, u^+ \in E$. This condition is satisfied for data in a sufficiently small neighborhood of a given point $\bar{u} \in \Omega$. We denote also with $\{l^i(u^-, u^+), i = 1, \dots, n\}$ the dual base.

We now define an approximated semigroup of solutions $\mathcal{S}^\nu$ on a set $E^\nu \subseteq E$. The construction is similar to the one in [3]. For any integer $\nu \in \mathbb{N}$, set

$$E^\nu \doteq \Big\{ u \in E : w_i(u) \in 2^{-\nu}\mathbb{Z}, i = 1, \dots, n \Big\}, \tag{2.10}$$

and let $D^{\nu,M}$ be the domain defined as

$$(2.11)\quad D^{\nu,M} \doteq \Big\{ u : \mathbb{R} \longmapsto E^\nu : u \text{ piecewise constant and Tot.Var.}(u) \leq M \Big\}.$$

Given $\bar{u} \in E^\nu$, we construct a solution $u(t)$ by wave front tracking. We first define how to solve the Riemann problem $[u^-, u^+]$, with $u^-, u^+ \in E^\nu$.

The solution to the Riemann problem u^-, u^+ is constructed by piecing together the solutions to the simple Riemann problems $[\omega_{i-1}, \omega_i]$, where ω_i is defined in (2.7). If the ith field is linearly degenerate, then $[\omega_{i-1}, \omega_i]$ is solved by a contact discontinuity traveling with speed $\lambda_i(\omega_i)$. If the ith field is genuinely nonlinear and $w_i^+ < w_i^-$, then $[\omega_{i-1}, \omega_i]$ is solved by a shock traveling with the Rankine–Hugoniot speed $\sigma_i(\omega_{i-1}, \omega_i)$. Finally, if the ith field is genuinely nonlinear and $w_i^+ > w_i^-$, then $[\omega_{i-1}, \omega_i]$ is solved by a rarefaction fan: if $w_i^+ = w_i^- + p_i 2^{-\nu}$, $p_i \in \mathbb{N}$, consider the states

$$\omega_{i,0} = \omega_{i-1}, \quad \omega_{i,l} = (w_1^+, \dots, w_{i-1}^+, w_i^- + \ell 2^{-\nu}, w_{i+1}^-, \dots, w_n^-), \quad \ell = 1, \dots, p_i.$$

The solution will consist of p_i shock waves $[\omega_{i,l-1}, \omega_{i,l}]$, traveling with the corresponding shock speed $\sigma_i(\omega_{i,l-1}, \omega_{i,l})$.

At time $t = 0$ we solve the initial Riemann problems of $\bar{u}$. Note that the number of wave fronts is bounded by $2^\nu \cdot \text{Tot.Var.}(\bar{u})$. When two or more fronts interact, we again solve the Riemann problem they generate, and so on. It is easy to show that at each interaction at least one of the following alternatives holds:

(i) the number of waves decreases at least by 1;

(ii) the total variation of the solution $u(t)$ decreases by $2^{1-\nu}$;

(iii) the interaction potential $Q(t)$, defined as

$$(2.12)\qquad Q(t) \doteq \sum_{\alpha,\beta \text{ approaching}} |\sigma_\alpha||\sigma_\beta| \leq M^2,$$

decreases by $2^{-\nu}$. We recall that two waves σ_α, σ_β of the families k_α, k_β, located at points x_α, X_β, are considered as *approaching* if $x_\alpha < x_\beta$ and $k_\alpha > k_\beta$.

This implies that there are at most a finite number of interactions, so that we can construct our approximate solution for all $t \geq 0$. Note that $\mathcal{S}_t^\nu u = u(t)$ is a semigroup of solutions, but not entropic due to the presence of rarefaction fronts.

If the ith family is linearly degenerate, the ith Riemann coordinate $w_i(t, \cdot)$ of the solution can be constructed by solving the semilinear system

$$(2.13)\qquad \begin{cases} (w_i)_t + \lambda_i(u(t,x))(w_i)_x = 0, \\ \qquad w_i(0,x) = w_{i,0}(x). \end{cases}$$

Since u is a piecewise constant solution, with a finite number of jumps, the broad solution to (2.13) is well defined [8]: If we denote with $x(t,y)$ the solution to the ODE

$$(2.14)\qquad \dot{x} = \lambda_i(u(t,x)), \qquad x(0) = y,$$

then the solution to (2.13) is given by

$$(2.15)\qquad w_i(t, x(t,y)) = w_{i,0}(y).$$

In the following sections we will consider the dependence on the initial data u_0 of the genuinely nonlinear Riemann coordinates $w_k(t, \cdot)$ and the map $h_i^t(y)$ defined as

$$(2.16)\qquad h_i^t(y) \doteq x_i(t,y),$$

where $x_i(t,y)$ is the solution to (2.14).

3. Estimates on the shift differential map. In this section we prove some properties of the shift differential map. These properties are closely related to the structure of (2.1), i.e., the conservation form, the coinciding shock and rarefaction assumption, which prevents the creation of shock when two jumps of the same family collide, and the existence of Riemann invariants, which prevents the creation of shock when two jumps of different families interact.

Consider a wave front solution $u(t, \cdot)$ of (2.1), and assume that the initial datum $u(0, \cdot)$ has a finite number N of jumps σ_α, located in y_α:

$$u(0,x) = \sum_{\alpha=1}^{N} \sigma_\alpha \chi_{[y_\alpha,+\infty)}(x).$$

If ξ_α is the shift rate of the jump σ_α, define $u^\theta(t, \cdot)$ as the front tracking solution with initial datum

$$u^\theta(0,x) = \sum_{\alpha=1}^{N} \sigma_\alpha \chi_{[y_\alpha+\theta\xi_\alpha,+\infty)}(x). \tag{3.1}$$

In the following, we will use the integral shift function, defined by

$$v(t,x) \doteq \lim_{\theta\to 0} \left\{ -\frac{1}{\theta} \int_{-\infty}^{x} u^\theta(t,y) - u(t,y)dy \right\}. \tag{3.2}$$

If $u(t, \cdot)$ has a shock σ_β, located in y_β, and if ξ_β is its shift rate, it is clear that the following relation holds:

$$\sigma_\beta \xi_\beta = v(t, y_\beta+) - v(t, y_\beta-). \tag{3.3}$$

We first recall the following result in [12], obtained using the conservation form of the equations.

LEMMA 3.1. *Consider a bounded, open region Γ in the t-x plane. Call σ_α, $\alpha = 1, \ldots, N$, the fronts entering Γ, and let ξ_α be their shifts. Assume that the fronts leaving Γ, say σ'_β, $\beta = 1, \ldots, N'$, are linearly independent. Then their shifts ξ'_β are uniquely determined by the linear relation*

$$\sum_{\beta=1}^{N'} \xi'_\beta \sigma'_\beta = \sum_{\alpha=1}^{N} \xi_\alpha \sigma_\alpha. \tag{3.4}$$

REMARK 3.2. *As observed in* [12], (3.4) *implies that the shift rates of the outgoing fronts depend only on the shift rates of the incoming ones, and not on the order in which these wave fronts interact inside Γ. In particular, we can perform the following operations, without changing the shift rates of the outgoing fronts:*

(O1) *switch the order of which three or more fronts interact;*

(O2) *invert the order of two fronts at time 0, if they have zero shift rate.*

The second lemma is concerned with a configuration where a sequence of contact discontinuities interacts with a wave of another family.

LEMMA 3.3. *Consider a family of parallel contact discontinuities σ_α of the ith linearly degenerate family, $\alpha = 1, \ldots, N$, and a single wave front σ of the kth family, $k \neq i$. Let ξ_α and ξ be their initial shifts, respectively, and let ξ'_α, ξ' be their shifts*

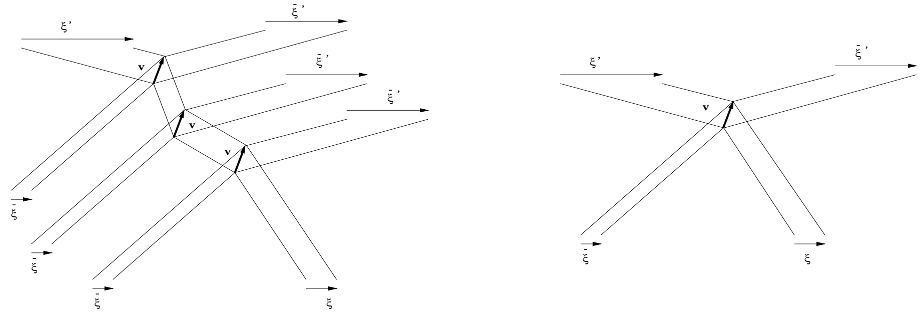


FIG. 2. *Interaction with a sheaf of contact discontinuities.*

after interaction. Assume that $\xi_\alpha = \bar\xi$ for all α. Then after the interactions all the shift rates ξ'_α of the ith family have the same value $\bar\xi'$ and

$$\xi'_\alpha = \bar\xi' = \frac{\bar\xi(\bar\Lambda' - \Lambda) - \xi(\bar\Lambda' - \bar\Lambda)}{\Lambda - \bar\Lambda}, \qquad \xi' = \frac{\bar\xi(\Lambda' - \Lambda) - \xi(\bar\Lambda' - \bar\Lambda)}{\Lambda - \bar\Lambda}, \tag{3.5}$$

where $\bar\Lambda$, Λ and $\bar\Lambda'$, Λ' are the speeds of the shocks σ_α and σ before and after interaction, respectively.

Proof. Define the vector $\mathbf{v}$ in the t-x plane as the shift of the first collision point. By a direct computation one finds

$$\mathbf{v} = \left(\frac{\xi - \bar\xi}{\bar\Lambda - \Lambda}, \frac{\bar\Lambda\xi - \Lambda\bar\xi}{\bar\Lambda - \Lambda} \right). \tag{3.6}$$

Since all the incoming shocks of the linearly degenerate family have the same speed $\bar\Lambda$, by simple geometrical considerations it follows that the vector $\mathbf{v}$ is constant during all interactions (Figure 2). Formula (3.5) follows easily. □

REMARK 3.4. *Note that this lemma allows us to perform the following new operation, without changing the shift rates:*

(O3) *replace a family of contact discontinuities σ_α of a linearly degenerate family, all with the same shift rate $\bar\xi$, by a single wave $\sigma = \sum \sigma_\alpha$ with shift rate $\bar\xi$.*

In the next lemma we will show that the existence of Riemann coordinates w implies a strong relation among shocks of different families.

LEMMA 3.5. *Consider two adjacent jumps belonging to different families, σ_i and σ_j, $i < j$, located at $x_i > x_j$. Let σ'_i, σ'_j be their strength after interaction. Then the following holds:*

$$\operatorname{span}\{\sigma_i, \sigma_j\} = \operatorname{span}\{\sigma'_i, \sigma'_j\}. \tag{3.7}$$

Proof. If ξ_i, ξ_j are the shift rates before interaction, and ξ'_i, ξ'_j after interaction, then (3.7) follows easily from the conservation relation

$$\sigma_i\xi_i + \sigma_j\xi_j = \sigma'_i\xi'_i + \sigma'_j\xi'_j \qquad \text{for all } \xi_i, \xi_j \in \mathbb{R}, \tag{3.8}$$

because, by assumption, no waves of other families are generated. □

REMARK 3.6. *Note that the previous lemma implies that the conservation relation (3.8) is bidimensional, i.e., the shocks σ_i, σ_j and σ'_i, σ'_j lie on a two dimensional*

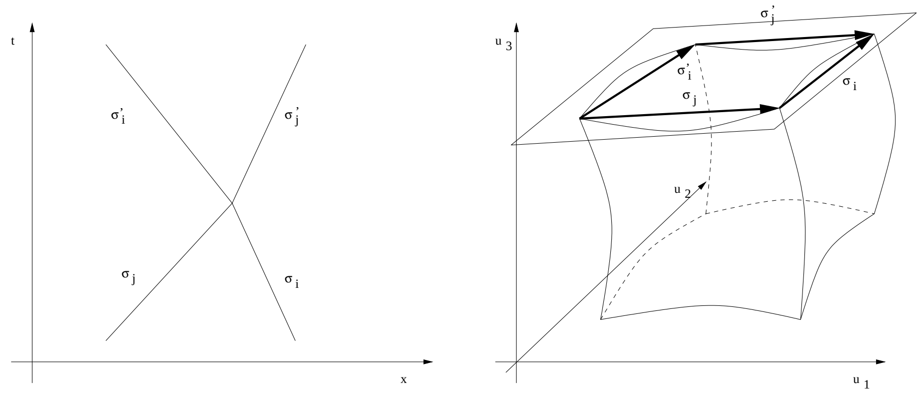


FIG. 3. *Vector relations among shocks.*

plane (Figure 3*). We can then obtain an identity which relates the strengths σ with the speeds Λ: substituting* (3.5) *into* (3.8)*, since $\bar{\xi}$, ξ are arbitrary, we get*

$$\begin{aligned} \sigma_i(\Lambda_j - \Lambda_i) &= \sigma'_i(\Lambda'_i - \Lambda_j) + \sigma'_j(\Lambda'_j - \Lambda_j), \\ \sigma_j(\Lambda_i - \Lambda_j) &= \sigma'_i(\Lambda'_i - \Lambda_i) + \sigma'_j(\Lambda'_j - \Lambda_i). \end{aligned} \tag{3.9}$$

One can show that if a Riemann solver verifies (3.9) *for all pairs of waves i, j, then there exists a flux function f such that the wave front approximation is a weak solution to* (2.1).

An important property of the shift differential map for Temple class systems is the fact that a perturbation to the initial data, initially localized in $[a, b]$, remains in the neighborhood of the set $\cup_i[x_i(t, a), x_i(t, b)]$, where $x_i(t, y)$ is the solution of the ith characteristic equation starting at y. We now extend this property to hyperbolic systems satisfying the hypotheses (H1), (H2), (H3) of section 2.

Consider N jumps σ_α, $\alpha = 1, \dots, N$, of some linearly degenerate family i, located at x_α and corresponding to the jumps $c(\alpha)e_i$ in the Riemann coordinates w:

$$\sigma_\alpha = u(w(x_\alpha-) + c(\alpha)e_i) - u(w(x_\alpha-)) \tag{3.10}$$

for some constants $c(\alpha)$, $\alpha = 1, \dots, N$.

DEFINITION 3.7. *We say that the jumps σ_α defined in* (3.10) *are in involution if*

$$\sum_{\alpha=1}^{N} c(\alpha) = 0, \tag{3.11}$$

i.e., the initial and final Riemann coordinate w_i is the same: $w_i(x_1-) = w_i(x_N+)$.

Note that, by the existence of Riemann coordinates, this relation does not depend on the positions and strength of the shocks of the other families. We can now extend Lemma 2 in [12] to our systems.

LEMMA 3.8. *Consider a wave front tracking solution u. Assume that there are N shocks σ_α either*

(i) *of the ith linearly degenerate family in involution, or*

(ii) *of the kth genuinely nonlinear family,*

and let $x_\alpha(t)$, $0 \leq t \leq T$, be the position of the shock σ_α, $\alpha = 1, \dots, N$. Then it is possible to assign at time $t = 0$ shift rates to all shocks such that $\xi_1 = 1$ and the shift of all fronts outside the strip $\Gamma \doteq \{(t, x); t \in [0, T], x_1(t) \leq x \leq x_N(t)\}$ is zero.

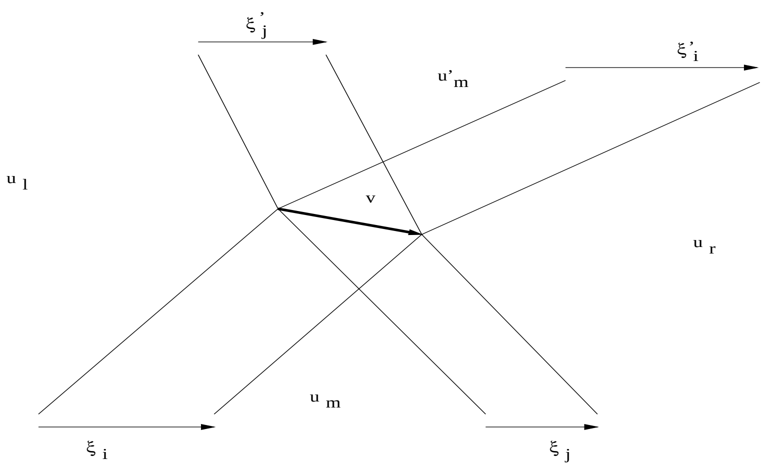


FIG. 4. *Computation of the shift rate.*

Proof. We consider only the case of linearly degenerate family i, since in the other case the proof is exactly the one given in [12].

Let $x_\alpha(t)$, $\alpha = 1, \dots, N$, be the position of the shock σ_α of the ith family in involution, and let $\bar{w}_i$ be the value of the Riemann coordinate at $x_1(t)- = x_1(0)-$. For $w \in E$, define $\tilde{w}$ as the projection of w on the hyperplane $\{w_i = \bar{w}_i\}$, and $\tilde{u} = u(\tilde{w})$.

We choose the shift rates such that

$$-\frac{d}{d\theta} \int_{-\infty}^{x} u^\theta dy = \sum_{x_i(t) \leq x} \xi_i(t)\sigma_i(t) = c(t,x)(u(t,x) - \tilde{u}(t,x)), \tag{3.12}$$

where $c(t,x)$ is a scalar function different from 0 only in $[x_1(t), x_N(t)]$, and we recall that $\tilde{u}(t,x) = u(\tilde{w}(t,x))$.

By imposing the value $\xi_1 = 1$, i.e., $c(0, x_1(0)-) = 0$, $c(0, x_1(0)+) = 1$, we need to prove that (3.12) can be satisfied at time $t = 0$. We have two cases.

(1) If the jump σ_i belongs to the ith family and is inside $[x_0(0), x_N(0)]$, then set $\xi = c(t, x_i-)$.

(2) If the jump σ_i belong to the kth family with $k \neq i$, then by assumption (2.8) and by (3.7) there exists a unique shift ξ_i and a unique constant $c(0,x+)$ such that

$$\xi_i\sigma_i + c(0,x-)(u(0,x-) - \tilde{u}(0,x-)) = c(0,x+)(u(0,x+) - \tilde{u}(0,x)).$$

Since we assume that the shocks are in involution, setting $\xi_N = c(0, x_n-)$ we have that (3.12) holds at time $t = 0$: in fact, the last jump has size $\tilde{u}(0, x_N(0)-) - u(0, x_N(0)-)$.

We now show that this property is conserved for all $t \geq 0$. This follows easily from conservation and Lemma 3.5. The proof is exactly the same as in [12]; we repeat it for completeness. Consider the interaction between two shocks σ_i and σ_j in the point (τ, y); see Figure 4. By inductive assumption, we have for the states u_l, u_m, and u_l that

$$\sum_{x_\gamma(\tau) < y} \sigma_\gamma(\tau)\xi_\gamma(\tau) = c_l(u_l - \tilde{u}_l), \tag{3.13}$$

$$c_l(u_l - \tilde{u}_l) + \sigma_i \xi_i = c_m(u_m - \tilde{u}_m),$$
$$c_m(u_m - \tilde{u}_m) + \sigma_j \xi_j = c_r(u_r - \tilde{u}_r).$$

Using conservation we have

$$\xi_i \sigma_i + \xi_j \sigma_j = \xi_j' \sigma_j' + \xi_i' \sigma_i', \tag{3.14}$$

so that for the new middle state u_m' we have

$$c_l(u_l - \tilde{u}_l) + \sigma_j' \xi_j' = c_m'(u_m' - \tilde{u}_m') = c_r(u_r - \tilde{u}_r) - \sigma_i' \xi_i', \tag{3.15}$$

and using Lemma 3.5 we conclude

$$\operatorname{span}\Big\{u_l - \tilde{u}_l, \sigma_j'\Big\} \bigcap \operatorname{span}\Big\{u_r - \tilde{u}_r, \sigma_i'\Big\} = \operatorname{span}\{u_m - \tilde{u}_m'\}.$$

The same relation proves that they vanish outside Γ. In fact, assume for example that $c_l = 0$ and $j < i$. Then from (3.15) we get

$$\sigma_j' \xi_j' = c_m'(u_m' - \tilde{u}_m'),$$

which implies that $c_m' = 0$. This concludes the proof. □

REMARK 3.9. *Note that for discontinuities of a linearly degenerate family all shift rates have the same sign. Note, moreover, that if no waves of other families are present, then we shift all jumps σ_α by unit rate* 1. *This corresponds to the case considered in Lemma* 3.3*, i.e., to the substitution of a family of contact discontinuities with a single jump, whose strength in this case is* 0 *by the involution assumption.*

Using conservation and the previous lemmas, we obtain explicitly the shift differential map at a given time τ. We recall that, given the states $u^-, u^+ \in E$, we denote with $r_i(u^-, u^+)$ the vectors defined in (2.9), and with $l^i(u^-, u^+)$ its dual base. Let $P_j(u^-, u^+)$ be the projection operator on $\operatorname{span}\{r_i(u^-, u^+), i = 1, \dots, j\}$:

$$P_j(u^-, u^+) v \doteq \sum_{i=1}^{j} \langle l^i(u^-, u^+), v \rangle r_i(u^-, u^+), \tag{3.16}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product in $\mathbb{R}^n$.

Given a point (t, x), with $u(t, x)$ continuous in x, define x_i to be the intersection of the backward ith characteristics starting at (t, x) with the real axis $\{(0, x)\}$, and for all $(0, y)$ let $j(y)$ be the index such that $x_{j(y)} \le y < x_{j(y)-1}$, $j(y) = 1, \dots, n+1$, with $x_0 = +\infty$ and $x_{n+1} = -\infty$. Without any loss of generality, we can assume that in $(0, y)$ there is a jump σ of the kth family.

Define the points $w_l, w_r \in E$ by

$$\begin{aligned} &w_l(x,y) \\ &\doteq \begin{cases} w(t,x), & j(y) = 1, \\ (w_1(0,y-), \dots, w_{j(y)-1}(0,y-), w_{j(y)}(t,x), \dots, w_n(t,x)), & 2 \le j(y) \le n, \\ w(0,y-), & j(y) = n+1, \end{cases} \\ &w_r(x,y) \\ &\doteq \begin{cases} w(0,y+), & j(y) = 1, \\ (w_1(t,x), \dots, w_{j(y)-1}(t,x), w_{j(y)}(0,y+), \dots, w_n(0,y+)), & 2 \le j(y) \le n, \\ w(t,x), & j(y) = n+1. \end{cases} \end{aligned} \tag{3.17}$$

Moreover, for $2 \leq j(y) \leq n$, define the point $w_m \in E$ by

$$\begin{aligned}(3.18)\quad & w_m(x,y) \\ & \doteq (w_1(t,x), \dots, w_k(0,y+), \dots, w_{j(y)-1}(t,x), w_{j(y)}(0,y+), \dots, w_n(0,y+))\end{aligned}$$

if $k < j(y)$. In a similar way, if $k \geq j(y)$,

$$\begin{aligned}(3.19)\quad & w_m(x,y) \\ & \doteq (w_1(t,x), \dots, w_{j(y)-1}(t,x), w_{j(y)}(0,y+), \dots, w_k(0,y-), \dots, w_n(0,y+)).\end{aligned}$$

Define $P(x,y)$ as the vector

$$(3.20)\qquad P(x,y) \doteq \begin{cases} 0, & j(y) = 1, \\ P_{j(y)-1}(w_l, w_m)\sigma + P_{j(y)-1}(w_m, w_r)\big(\sigma - P_{j(y)-1}(w_l, w_m)\sigma\big), & \\ \qquad 2 \leq j(y) \leq n+1, k < j(y), & \\ P_{j(y)-1}(w_l, w_m)\big(P_{j(y)-1}(w_m, w_r)\sigma\big), & 2 \leq j(y) \leq n, k \geq j(y), \end{cases}$$

where $w_l = w_l(x,y)$, $w_m = w_m(x,y)$, $w_r = w_r(x,y)$, and σ is the initial jump in $(0,y)$. Consider now a front tracking solution u^θ, obtained by shifting the initial jumps σ_α in y_α with rates ξ_α.

THEOREM 3.10. *If $v(t,x)$ is the integral shift function of $u^\theta(t,\cdot)$, defined in* (3.2), *then*

$$(3.21)\qquad v(t,x) = \lim_{\theta \to 0} \left\{ -\frac{1}{\theta} \int_{-\infty}^{x} u^\theta(t,y) - u(t,y) dy \right\} = \sum_\alpha P(x, y_\alpha)\xi_\alpha.$$

Proof. The theorem will be proved outside the times of interaction, because the Lipschitz dependence in L^1 of the approximate semigroup implies the validity of (3.21) for all $t \geq 0$.

If is sufficient to show that $\sum_{y_\alpha} P(x, y_\alpha)\xi_\alpha$ is piecewise constant, with jumps only at the points x_β where $u(t,\cdot)$ has a shock σ_β, and the following relation holds:

$$(3.22)\ \sum_{y_\alpha} \big(P(x_\beta+, y_\alpha) - P(x_\beta-, y_\alpha)\big)\xi_\alpha = \sigma_\beta \xi_\beta, \qquad \lim_{x \to -\infty} \sum_{y_\alpha} P(x, y_\alpha)\xi_\alpha = 0,$$

where ξ_β is the shift rate of σ_β, located in x. Note that by (3.20) the second equality of (3.22) is trivially satisfied.

By linearity in the shift rates ξ_α, we can consider the case in which a single shock is shifted, let us say σ at y; (3.21) becomes

$$(3.23)\qquad v(t,x) = P(x,y)\xi.$$

Formula (3.16) follows from the following considerations: Consider a wave front pattern, Figure 5, where for simplicity we assume that $k < j(y)$. The states w_l, w_m are computed considering the Riemann problem generated by adding to the k-jump σ in $(0,y)$ all the i-waves starting from the left of $(0,y)$ and ending in the right of (t,x) and all the i-waves, with $i \neq k$, starting from the right of $(0,y)$ and ending in the left of (t,x). The jump w_m, w_r is a single wave of the kth family formed by adding all the k-waves between $(0,y)$ and (t,x). Using the definition of $v(t,x)$ given (3.2), one obtains easily the second case of (3.20). In fact

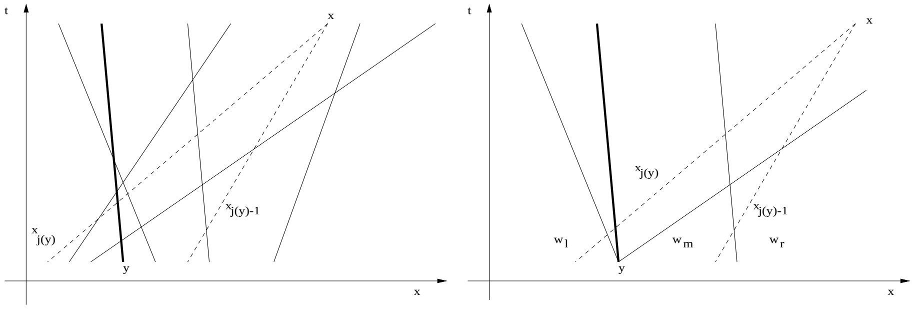


FIG. 5. *Wave pattern for the computation of formula* (3.16).

the shift rates of the shocks in the left of (t, x) are given by the shift rates of the jumps of the Riemann problem w_l, w_m ending in the left of (t, x), $P_{j(y)-1}(w_l, w_m)\sigma$, plus the shift rate of the shock w_m, w_r, $P_k(w_m, w_r)(\sigma - P_{j(y)-1}(w_l, w_m)\sigma)$. Since only the i-waves with $i \geq j(y) > k$ are present in $\sigma - P_{j(y)-1}(w_l, w_m)\sigma$, then $P_k(w_m, w_r)(\sigma - P_{j(y)-1}(w_l, w_m)\sigma) = P_{j(y)-1}(w_m, w_r)(\sigma - P_{j(y)-1}(w_l, w_m)\sigma)$. The other cases can be computed in a similar way: in this case one solves the Riemann problem w_m, w_r in $(0, y)$ and considers the k-wave w_l, w_m starting in the left of $(0, y)$ and ending in the right of (t, x).

From the above considerations, it is clear that $P(x, y)$ is piecewise constant with jumps only when there is an i-shock σ' in (t, x). In fact, otherwise, the wave front pattern used to compute $P(x, y)$ remains the same. Let $\{z_p : p = 1, \dots, M\}$ be the set of the starting points of all shocks arriving in (t, x), and define

$$z^- = \min_p z_p, \qquad z^+ = \max_p z_p. \tag{3.24}$$

We consider two cases:

(1) The shocks arriving in (t, x) start on both sides of $(0, y)$: $z^- \leq y \leq z^+$. In this case, $(P(x+, y) - P(x-, y))\xi$ is the shift rate of the i-shock starting in the Riemann problem $w_l(x-, y), w_m(x+, y)$ if $i > k$ ($w_m(x-, y), w_r(x+, y)$ if $i < k$), which collides with a k-shock $w_m(x+, y), w_r(x+, y)$ (w_l, w_m if $i < k$). In fact, the only difference is that in $w_m(x-, y), w_m(x+, y)$ there is a shock of the ith family starting in $(0, y)$, and i is genuinely nonlinear. Finally, using $r_i \bullet r_i(u) = 0$ and Lemma 3.5, one can change position to the i-wave and the remaining k-wave w_m, w_r, whose strength does not change.
If $i = k$, there are no k-shocks starting on the right (left) of $(0, y)$ and ending on the right (left) of (t, x), so that $(P(x+, y) - P(x-, y))\xi$ is the shift rate of the i-shock of the Riemann problem $w_l(x-, y), w_r(x+, y)$.

(2) The shocks of the ith family arriving in (t, x) start either in $(-\infty, y)$ or $(y, +\infty)$. Assume for definiteness that $y < z^-$. In this case the difference $(P(x+, y) - P(x-, y))\xi$ is the shift rate of the shock σ' colliding with the shifted shocks of the Riemann problem $w_l(x-, y), w_m(x-, y)$ in $(0, y)$, crossing the jump $w_m(x-, y), w_r(x-, y)$, and finally overtaking σ'. In fact one can use Lemma 3.5 (and $r_i \bullet r_i(u) = 0$ if i is genuinely nonlinear) to obtain the wave pattern of Figure 6.

The various cases will be proved in the following lemmas.

LEMMA 3.11. *Assume that $z^- \leq y \leq z^+$, i.e., case* (1). *If the shock σ' is of the*

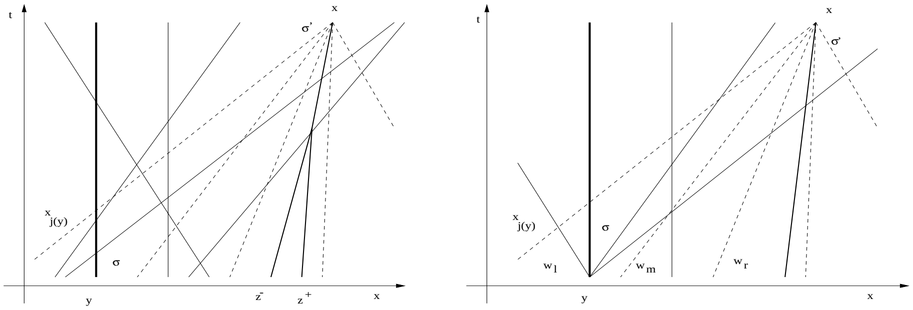


FIG. 6. *Computation of the shift rate in the case of Lemma* 3.12.

ith family, then its shift ξ' is

$$\xi'\sigma' = \big(P(x+,y) - P(x-,y)\big)\xi. \tag{3.25}$$

Proof. We follow closely the method of [12]. Assume for definiteness $k < j(y)$, the other cases being similar. The basic idea is to reduce the computation to the single Riemann problem $w_l(x-,y), w_m(x+,y)$, with eventually a single k-wave w_m, w_r.

Consider Figure 7. By Lemma 3.1, we can simplify the wave configuration considering only the fronts crossing starting in the right of $(0,y)$ and ending in the left of (t,x). In fact we can move the other fronts to $\pm\infty$ without changing the shift rate of σ'.

We can now shift the initial position of the waves of the ith family merging in x such that their initial position coincides with y, without changing the shift rate ξ'. This operation can be repeated for all shocks of genuinely nonlinear families.

Finally, we can move the shocks of the linearly degenerate families such that they have the same sequence of interaction with the other shocks. This means that, if x_i^j is the position of the jth shock of the ith linearly degenerate family, the only interactions among shocks occurring in the sector $[x_i^1(t), x_i^n(t)]$ are those involving one ith wave and one kth wave, with $k \neq i$. Using Lemma 3.3, we can at this point substitute them with a single shock, whose strength is the sum of the strengths of the i-waves. Finally we move their position at $t = 0$ such that it coincides with y; we obtain the wave patterns of Figure 7. To conclude, we just need to prove that the Riemann problem obtained in this way is exactly $w_l(x-,y), w_m(x+,y)$ and that the remaining k-wave is $w_m(x+,y), w_r(x+,y)$.

By the previous argument, the strength of the shock of the jth family, $j < i$, $j \neq k$, is given by the j-waves starting in the right of y and ending in the left of x. Since they are the only j-wave crossing the segment $[(0,y-),(t,x+)]$, it follows that

$$w_{l,j}(x,y) = w_j(0,y-), \qquad w_{r,j}(x,y) = w_j(t,x+).$$

The other relations for $j = k$ and $j > i$ follow in the same way. Finally, for $j = i$ the jump is $w_i(t,x+) - w_i(t,x-)$. Note that the wave pattern is the same as that obtained in case (1). □

We consider only the case $y < z^-$, since the other is entirely similar.

LEMMA 3.12. *Assume that $y < z^-$. Then the shift ξ' of σ' is given by*

$$\xi'\sigma' = \big(P(x+,y) - P(x-,y)\big)\xi. \tag{3.26}$$

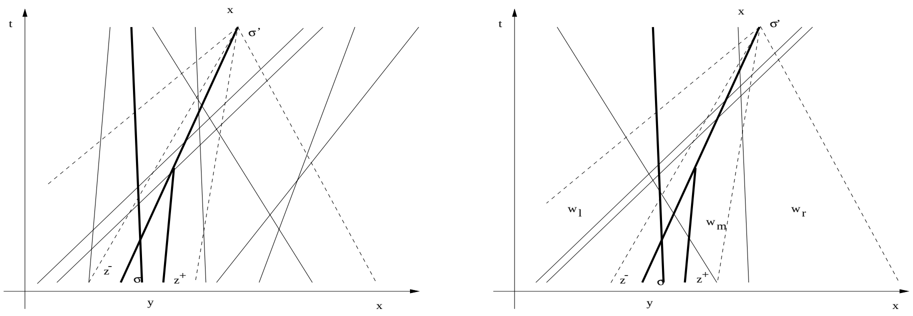


FIG. 7. *Computation of the shift rate in the case of Lemma* 3.11.

Proof. The hypothesis implies that the ith shock ending at x starts in the right of y. With the same simplification considered in Lemma 3.11, we reduce to the Riemann problem $w_l(x,y), w_r(x,y)$ in $\bar{y}$, such that the waves of the jth families, $j > i$, generated at $\bar{y}$ collide with the i-wave in x_α (see Figure 6) after overtaking the k-wave w_m, w_r. The conclusion follows easily, since the wave pattern is the same as that considered in (2). □

This concludes the proof of Theorem 3.10. □

Finally we extend to our case the following result proved in [12].

PROPOSITION 3.13. *Let u be a wave front tracking solution, and consider two wave fronts, $x(t)$ and $y(t)$, $t \in [0,T]$. Then there exists a second front tracking solution $\tilde{u}$ such that the initial and final positions of the two shocks are the same, and* Tot.Var.$(\tilde{u})$ *is uniformly bounded.*

Proof. For genuinely nonlinear fields, the proof is the same as in [12]. We then restrict the proof to the case of linearly degenerate fields i.

Assume that there exist two jumps σ_1, σ_2 of the ith family, with positions $z_1(t) < z_2(t)$, such that

$$\begin{aligned} x(0) \notin [z_1(0), z_2(0)] \text{ and } y(0) \notin [z_1(0), z_2(0)], \\ x(T) \notin [z_1(T), z_2(T)] \text{ and } y(T) \notin [z_1(T), z_2(T)]. \end{aligned} \tag{3.27}$$

For definiteness, assume $w_i(0, z_1-) < w_i(0, z_1+)$, and the following condition is satisfied:

$$w_i(0, z_1-) \in [w_i(0, z_2-), w(0, z_2+)]. \tag{3.28}$$

Let σ_α, $\alpha = 1, \dots, N$, be the jumps of linearly degenerate family i in the strip $[z_1(0), z_2(0))$. If we define

$$\sigma_{N+1} = u(w_i(0, z_1-)) - u(w_i(0, z_2-)),$$

it is easy to verify that the shocks σ_α, $\alpha = 1, \dots, N+1$, are in involution. By Lemma 3.8, we can then move the jumps to the left until either $z_1(t)$ meets the wave fronts $x(t)$, or $z_1(t)$ coincides with another shock of the ith family (Figure 8). It is clear that we can repeat the same procedure also in the following cases:

(i) $w_i(0, z_1-) > w_i(0, z_1+)$ and $w_i(0, z_1-) \in [w_i(0, z_2+), w(0, z_2-)]$;

(ii) $w_i(0, z_2-) < w_i(0, z_2+)$ and $w_i(0, z_2+) \in [w_i(0, z_1+), w(0, z_1-)]$;

(iii) $w_i(0, z_2-) > w_i(0, z_2+)$ and $w_i(0, z_2+) \in [w_i(0, z_1-), w(0, z_1+)]$.

It is now easy to prove that the total variation of the jumps of the ith family satisfying (3.27) can be at most $3\|w\|_\infty$. Since $x(t)$, $y(t)$ divide the lines $t = 0$ and $t = \tau$ in three regions, the total variation of w_i is bounded by $27\|w\|_\infty$. □

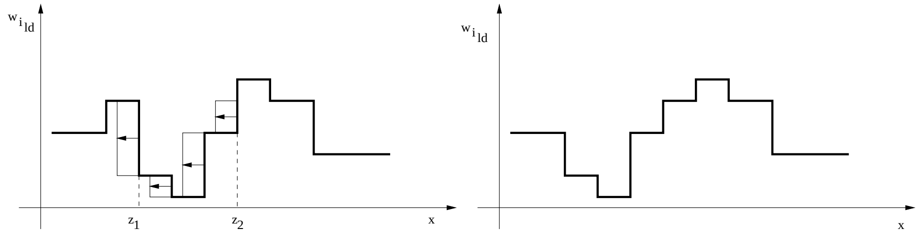


FIG. 8. *Cancellation among contact discontinuities.*

4. Estimates on characteristics. In this section we prove some estimates on the solution $x_i(t,y)$ of the characteristic equation

$$\begin{cases} \dot{x}_i = \lambda_i(u(t,x_i)), \\ x_i(0) = y. \end{cases} \tag{4.1}$$

We assume for simplicity that the ith family is linearly degenerate; however, the same results are valid for characteristics of a genuinely nonlinear family if the following condition holds: for all τ there exists an ϵ such that in the strip $\{(t,x); \tau \leq t \leq T, x_i(t,y) - \epsilon \leq x \leq x_i(t,y) + \epsilon\}$ there are no shock waves of the ith family. Given front tracking, approximation u, $x_i(t,y)$ is unique, since it crosses only a finite number of transversal jumps, and it depends Lipschitz continuously on the initial data y (see [8]).

We want to give uniform estimates on this dependence. The idea is to suppose that in y there is a shock σ^ϵ of the i-family of size ϵ: $w_i(0,y+) - w_i(0,y-) = \epsilon$. Since by assumption no shocks of the i-family collide with σ^ϵ, it is easy to construct a wave front solution. For $x < x(t,y)$, the solution $u^\epsilon(t,\cdot)$ takes values in

$$E^{\nu,-} \doteq \Big\{u : w_j(u) \in [a_j, b_j] \cap 2^{-\nu}\mathbb{Z}, j = 1,\ldots,n\Big\},$$

while for $x > x(t,y)$, enlarging E and assuming ϵ sufficiently small,

$$E^{\nu,+} \doteq \Big\{u : w_j(u) \in [a_j,b_j] \cap 2^{-\nu}\mathbb{Z}, j \neq i, w_i(u) \in [a_i,b_i] \cap \big\{2^{-\nu}\mathbb{Z} + \epsilon\big\}\Big\}.$$

The following lemma proves the continuous dependence of the solution $u^\epsilon(t)$ and the position $x_i^\epsilon(t,y)$ of the shock σ^ϵ w.r.t. ϵ.

LEMMA 4.1. *Consider a front tracking solution u, with initial data u_0 and the characteristic lines $x_i(t,y_1) < x_i(t,y_2)$, defined in* (4.1) *for a linearly degenerate family i. Let u^ϵ be the wave front solution with initial data $u(w_0^\epsilon)$, where w_0^ϵ is defined as*

$$w_0^\epsilon(x) \doteq \begin{cases} w(u_0(x)), & x \leq y_1, \\ w(u_0(x)) + \epsilon e_i, & y_1 < x \leq y_2, \\ w(u_0(x)), & x > y_2. \end{cases} \tag{4.2}$$

Then there exist constants L, L', depending only on the total variation of the initial data u_0, such that for all $t \geq 0$

$$\begin{aligned} &\int_{\mathbb{R}} \big|u(t,x) - u^\epsilon(t,x)\big| dx \leq L\epsilon\big|y_1 - y_2\big| \quad \textit{and} \\ &\big|x_i^\epsilon(t,y_j) - x_i(t,y_j)\big| \leq L'\epsilon t\big|y_1 - y_2\big|, \quad j = 1,2, \end{aligned} \tag{4.3}$$

where $x_i^\epsilon(t, y_j)$ is the position of the shock σ_j^ϵ starting in $(0, y_j)$.

Proof. The first inequality is an easy consequence of the L^1 continuous dependence for front tracking solutions; see [3]. For the second one, note that all the shocks different from σ^ϵ have size uniformly bigger than 0, so that their position is shifted of the order ϵ. Thus the second inequality follows by standard ODE perturbation estimates; see [8]. □

An easy application of the previous lemma together with Proposition 3.13 implies that to compute $x_1(t, y_1)$ and $x_2(t, y_2)$ we can actually consider in (4.1) a solution $\tilde{u}$ with uniformly bounded total variation, so that the constant L' in (4.3) is independent on the total variation of u_0.

We now estimate the dependence of $x_i(t, y)$ w.r.t. u.

PROPOSITION 4.2. *Let ξ_α be the shift rate of the jump σ_α in $u(0, \cdot)$, and denote with x_i^θ the solution to*

$$\begin{cases} \dot{x}_i^\theta = \lambda_i(u^\theta(t, x_i^\theta)), \\ x_i^\theta(0) = y, \end{cases}$$

where $u^\theta(t)$ is the shifted front tracking solution. Then there exists a constant D independent of the total variation of u such that

$$\left|\lim_{\theta \to 0} \frac{x_i^\theta(t, y) - x_i(t, y)}{\theta}\right| \leq D \sum_\alpha \Big|\sigma_\alpha \xi_\alpha\Big|. \tag{4.4}$$

Proof. If ϵ is the size of the shock σ^ϵ located in $(0, y)$, then we can apply Theorem 3.10 to compute its shift ξ^ϵ. By formula (3.21) we obtain

$$\xi^\epsilon \sigma^\epsilon = \sum_\alpha \big(P(x+, y_\alpha) - P(x-, y_\alpha)\big)\xi_\alpha. \tag{4.5}$$

If θ is sufficiently small, then we have

$$\xi^\epsilon = \frac{x_i^{\theta,\epsilon}(t, y) - x_i^\epsilon(t, y)}{\theta},$$

where $x_i^{\theta,\epsilon}(t, y)$ is the position of the shifted shock and $x_i^\epsilon(t, y)$ is its original position. Note that $(P(x+, y_\alpha) - P(x-, y_\alpha))\xi_\alpha$ is the shift rate of the shock σ^ϵ after colliding with the shocks of the Riemann problems w_l, w_m and w_m, w_r. Their total shift is proportional to $|\sigma_\alpha \xi_\alpha|$, and after the interaction with σ^ϵ, the shift of the latter is proportional to $|\sigma^\epsilon||\sigma_\alpha \xi_\alpha|$. Thus taking the limit as ϵ tends to 0 of (4.5), we obtain for ϵ sufficiently small

$$\left|\frac{x_i^\theta(t, y) - x_i(t, y)}{\theta}\right| \leq D \sum_\alpha \Big|\sigma_\alpha \xi_\alpha\Big|,$$

which implies (4.4). □

We now prove the uniform Lipschitz continuity of the map $y \longmapsto x_i(t, y)$ for all $t \geq 0$.

PROPOSITION 4.3. *Consider two characteristic lines $x_i(t, y_1)$, $x_i(t, y_2)$, solutions to (4.1). There exists $C > 0$, depending only on the system and the set E, such that*

$$\frac{1}{C} \leq \frac{x_i^2(t, y_2) - x_i^1(t, y_1)}{y_2 - y_1} \leq C. \tag{4.6}$$

Proof. As in the previous proposition, let ϵ be the size of the shock $\sigma^\epsilon(t)$ located in $(0,y)$ in Riemann coordinates. If $\xi(t)$ is its shift rate, then for θ sufficiently small by Theorem 3.10 we obtain

$$\frac{x_i^\epsilon(t,y+\theta\xi)-x_i^\epsilon(t,y)}{\theta}\sigma^\epsilon(t)=\xi^\epsilon(t)\sigma^\epsilon(t)=r_i(w_l,w_r)\bigl\langle l^i(w_l,w_r),\sigma(0)\bigr\rangle\xi(0). \tag{4.7}$$

In fact, by assumption, in the simplified wave patterns to compute the shift rate of σ^ϵ, there are no waves of the ith family different from σ^ϵ. Dividing by ϵ and taking the limit as ϵ tends to 0, we obtain

$$\frac{x_i(t,y+\theta\xi)-x_i(t,y)}{\theta}\frac{\partial}{\partial w_i}u(t,x)=\frac{\partial}{\partial w_i}u(0,y)\bigl\langle l^i(w_l,w_r),r_i(0,y)\bigr\rangle\xi(0),$$

which implies

$$\frac{d}{dy}x_i(t,y)=\frac{\partial u(0,y)/\partial w_i}{\partial u(t,x)/\partial w_i}\bigl\langle l^i(w_l,w_r),r_i(0,y)\bigr\rangle. \tag{4.8}$$

We use the fact that $\sigma^\epsilon(t)/\epsilon$ tends to $\partial u(0,y)/\partial w_i\cdot r_i(w_l,w_r)$ as $\epsilon\to 0$. Since E is compact, the conclusion (4.6) follows easily. $\square$

Remark 4.4. *The above proposition implies that the map h_i^t defined in* (2.16) *is uniformly Lipschitz, independent on the total variation of u_0, together with its inverse map $(h_i^t)^{-1}$.*

To end this section, we give a different proof of the following result given in [12].

Proposition 4.5. *If $x(t)$, $y(t)$ are the positions of two adjacent k-rarefaction waves, then for some constant $\kappa>0$ one has*

$$y(\tau)-x(\tau)\geq\kappa\tau 2^{-\nu}, \tag{4.9}$$

where $c>0$ is the constant defined in (2.3). *Thus for all $\tau>0$ the total variation of the Riemann invariant w_k of the kth genuinely nonlinear family with N shocks at $t=0$ is bounded by*

$$\text{Tot.Var.}\{w_k(\tau,\cdot);[a,b]\}\leq\frac{2(b-a)}{\kappa\tau}+\bigl\|w_k\bigr\|_{L^\infty}+(N+1)2^{1-\nu}. \tag{4.10}$$

Proof. Consider two adjacent k-rarefaction fronts $x(t)$ and $y(t)$, and let t_α, $\alpha=1,\dots,N$, be the interaction times of $x(t)$, $y(t)$ with other waves in the interval $[0,\tau]$. Fix $t_i\in(t_{\bar\alpha},t_{\bar\alpha+1})$ for some $\bar\alpha$ and let $z(t,x(t_i))$ be the characteristic line of the kth genuinely nonlinear family starting in $(t_i,x(t_i))$ (see Figure 9). Assume $t_{i+1}>t_i$ sufficiently close to t_i such that $t_{i+1}\in(t_{\bar\alpha},t_{\bar\alpha+1})$ and $z(t,x(t_i))$ does not collide with shocks of other families for $t\in[t_i,t_{i+1}]$. Let $z(t,x(t_{i+1}))$ be the characteristic curve starting in $(t_{i+1},x(t_{i+1}))$. By the assumption of genuine nonlinearity, at time t_{i+1} we have

$$z(t_{i+1},x(t_i))-z(t_{i+1},x(t_{i+1}))\geq c(t_{i+1}-t_i)2^{-\nu-1}$$

for some constant c, depending only on E. Using Proposition 4.3, at time τ we have

$$z(\tau,x(t_i))-z(\tau,x(t_{i+1}))\geq\frac{c}{C}(t_{i+1}-t_i)2^{-\nu-1}. \tag{4.11}$$

Repeating the process, it is possible to find a countable number of times t_i such that

$$\lim_{i\to-\infty}t_i=t_{\bar\alpha},\qquad\lim_{i\to+\infty}t_i=t_{\bar\alpha+1},$$

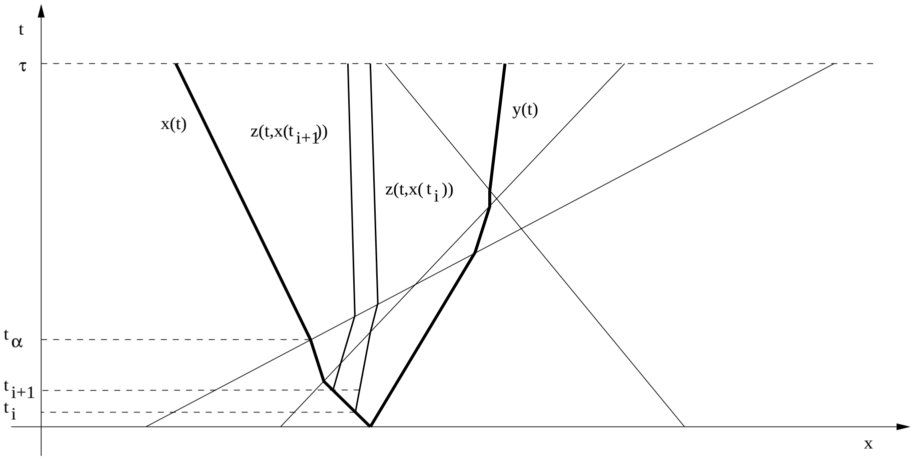


FIG. 9. *Decay of positive waves.*

and using (4.11) we get

$$z(\tau, x(t_{\bar\alpha})) - z(\tau, x(t_{\bar\alpha+1})) \geq \frac{c}{C}(t_{\bar\alpha+1} - t_{\bar\alpha})2^{-\nu-1}. \tag{4.12}$$

Repeating the process for $y(t)$ and for all intervals $(t_{\alpha+1}, t_\alpha)$, we obtain (4.9), where $\kappa = c/C$.

The second equation follows, noticing that the total amount of positive jumps in the interval $[a, b]$ is bounded by $(1 + N)2^{-\nu} + (b - a)/\kappa\tau$. □

5. Proof of the main theorem. In this section we construct the semigroup $\mathcal{S}$ on $L^\infty(\mathbb{R}; E)$. In [3] it is shown that for all M there exists a semigroup $\mathcal{S}^M$ defined on the domain

$$D^M \doteq \Big\{u : \mathbb{R} \mapsto E : \text{Tot.Var.}(u) \leq M\Big\}, \tag{5.1}$$

which is the only limit of the wave front tracking approximations constructed in section 2. We now study the dependence of the solution on the initial data $u \in D^M$. We consider separately the case for genuinely nonlinear and linearly degenerate families.

PROPOSITION 5.1. *Consider a front tracking solution u, such that $u(0, \cdot)$ has N jumps σ_α, $\alpha = 1, \ldots, N$, and let ξ_α be their shift rates. Given $\tau \geq 0$, denote with σ_β the jumps in the Riemann invariant $w_k(\tau, \cdot)$ of the kth genuinely nonlinear family. Then there exists a constant K, depending only on the system and the domain E, such that*

$$\sum_\beta \big|\xi_\beta \sigma_\beta\big| \leq K(1 + N2^{-\nu}) \sum_{\alpha=1}^N \big|\xi_\alpha \sigma_\alpha\big|. \tag{5.2}$$

Proof. The proof follows by Theorem 3.10 and Proposition 4.5. In fact, given a fixed shock $\sigma_{\bar\alpha}$, using Theorem 3.10, we have that at time τ for a shock σ_β of the ith family there exist D' such that

$$\big|\xi_\beta \sigma_\beta\big| \leq D' \big|\xi_{\bar\alpha} \sigma_{\bar\alpha}\big| \tag{5.3}$$

if the shock σ_β starts on both sides of $\sigma_{\bar\alpha}$, or, using the same estimate of Proposition 4.2,

$$|\xi_\beta| \leq D|\sigma_{\bar\alpha}\xi_{\bar\alpha}| \tag{5.4}$$

if σ_β starts on one side of $\sigma_{\bar\alpha}$. Since there is at most 1 shock such that (5.3) holds, and the interval of influence is $[x_{\bar\alpha} - \hat\lambda\tau, x_{\bar\alpha} + \hat\lambda\tau]$, using Proposition 4.5 together with (5.3) and (5.4) we obtain

$$\sum_\beta |\xi_\beta\sigma_\beta| \leq D'|\xi_{\bar\alpha}\sigma_{\bar\alpha}| + D|\sigma_{\bar\alpha}\xi_{\bar\alpha}| \cdot \text{Tot.Var.}\Big\{w_k, [x_{\bar\alpha} - \hat\lambda\tau, x_{\bar\alpha} + \hat\lambda\tau]\Big\} \leq F(1+2^{-\nu})|\xi_{\bar\alpha}\sigma_{\bar\alpha}|.$$

The conclusion follows the linearity of the shift differential map. □

Using the results of the previous section, the following result is trivial.

Proposition 5.2. *Consider a wave front solution u, such that $u(0,\cdot)$ has N jumps σ_α, $\alpha = 1,\dots,N$, and let ξ_α be their shifts. Consider* (4.1), *with the eigenvalue λ_i linearly degenerate. For fixed $\tau \geq 0$, the shift ξ_i of $x_i(\tau, y)$ is then bounded by*

$$|\xi_i| \leq D\sum_{\alpha=1}^N |\xi_\alpha\sigma_\alpha|. \tag{5.5}$$

Proof. This is a corollary of Proposition 4.2. □

Using the above propositions, we can prove the following theorem.

Theorem 5.3. *Consider two initial data u_1 and u_2, and denote with $w_{j,k}(t,\cdot)$ the kth Riemann coordinate of $\mathcal{S}^M u_j$, $j = 1,2$, corresponding to the kth genuinely nonlinear family. Moreover, let $h^\tau_{j,i}$, $j = 1,2$, be the map defined in* (4.1) *for the ith linearly degenerate family. Then there exists a constant K', independent of M, such that the following estimates hold:*

$$\int_{\mathbb{R}} |w_{1,k}(t,x) - w_{2,k}(t,x)|dx \leq K' \int_{\mathbb{R}} |u_1(x) - u_2(x)|dx, \tag{5.6}$$

$$\sup_{t\geq 0, x\in\mathbb{R}} |h^t_{1,i}(x) - h^t_{2,i}(x)| \leq K' \int_{\mathbb{R}} |u_1(x) - u_2(x)|dx. \tag{5.7}$$

Proof. Consider two piecewise constant initial data u^ν_1, u^ν_2 in $D^{M,\nu}$, and construct a pseudopolygonal path $\gamma_0 : \theta \longmapsto u^\nu_\theta$, connecting u_1 and u_2, such that

$$\|\gamma_0\|_{L^1} \leq E\|u^\nu_1 - u^\nu_2\|_{L^1}.$$

We can assume that u^ν_θ has a finite number N of jumps. If we denote with γ^ν_τ the path $\theta \longmapsto \mathcal{S}^\nu_\tau u^\nu_\theta$, we have by Proposition 5.1

$$\begin{aligned} \|w^\nu_{2,k}(\tau) - w^\nu_{1,k}(\tau)\|_{L^1} &\leq \left\|(\gamma^\nu_\tau)_k\right\|_{L^1} \leq K(1 + N2^{-\nu})\|\gamma_0\|_{L^1} \\ &\leq K'(1 + N2^{-\nu})\|u_2 - u_1\|_{L^1}. \end{aligned} \tag{5.8}$$

Now if $\nu \to +\infty$, since $w^\nu_{j,k}(\tau)$ converges to $w_{j,k}(\tau)$, we obtain (5.6). Since this estimate does not depend on the number of initial jumps N, we can extend it uniformly on D^M.

Using the same pseudopolygonal path, in a similar way we can prove that

$$\left|x^\nu_{2,i}(\tau,y)-x^\nu_{1,i}(\tau,y)\right|\le K'\left\|u_2-u_1\right\|_{L^1}.$$

This shows that $x^\nu_i(\tau,\cdot)$ converges uniformly to the solution $x_i(\tau,\cdot)$ as $\nu\to+\infty$ and $u^\nu\to u$. It also implies that

$$\left|x_{2,i}(\tau,y)-x_{1,i}(\tau,y)\right|\le K'\left\|u_2-u_1\right\|_{L^1}.$$

This concludes the proof. □

We can now define $\mathcal{S}$ on the domain $L^\infty(\mathbb{R};E)$.

Definition 5.4. *For all $u\in L^\infty(\mathbb{R},E)$, let $u^M\in D^M$ be such that*

$$\lim_{M\to+\infty}u^M=u\quad in\ L^1_{\mathrm{loc}}. \tag{5.9}$$

Define $\mathcal{S}_tu$ as

$$\mathcal{S}_tu=\lim_{M\to+\infty}\mathcal{S}^M_tu, \tag{5.10}$$

where the limit is in L^1_{loc}.

It is easy to prove that the right-hand side of (5.10) is a Cauchy sequence in every compact set $[a,b]$. In fact, using the finite speed of propagation, we can consider u with compact support $[a-\hat\lambda t,b+\hat\lambda t]$. For the components w_k of the kth genuinely nonlinear family, it follows directly from (5.6), while for a linearly degenerate component w_i, let $\tilde w$ be a Lipschitz continuous function such that

$$\int_{\mathbb{R}}\left|w_i(0,x)-\tilde w(x)\right|dx\le\epsilon.$$

By Theorem 5.3 we have for $u_1,u_2\in D^M$ such that $\|u-u_i\|_{L^1}<\delta$, $i=1,2$,

$$\sup_{t\ge0,x\in\mathbb{R}}\left|h^t_{1,i}(x)-h^t_{2,i}(x)\right|<K'\delta,$$

and it follows by easy computations that

$$\begin{aligned}\left\|w_{1,i}(t)-w_{2,i}(t)\right\|_{L^1}&\le\left\|w_{1,i}(t)-\tilde w\circ\left(h^t_{1,i}\right)^{-1}\right\|_{L^1}\\&\quad+\left\|w_{2,i}(t)-\tilde w\circ\left(h^t_{2,i}\right)^{-1}\right\|_{L^1}\\&\quad+\left\|\tilde w\circ\left(h^t_{1,i}\right)^{-1}-\tilde w\circ\left(h^t_{2,i}\right)^{-1}\right\|_{L^1}\\&\le C\left\|w_{1,i}(0)-\tilde w\right\|_{L^1}+C\left\|w_{2,i}(0)-\tilde w\right\|_{L^1}\\&\quad+L(b-a)G\left\|u_2-u_1\right\|_{L^1}\\&\le2C(\epsilon+\delta)+L(b-a)G\delta,\end{aligned} \tag{5.11}$$

where L is the Lipschitz constant of $\tilde w$. This shows that $w^M_i(t)$ is a Cauchy sequence for all $t\ge0$, because the right-hand side of (5.11) can be made arbitrarily small. We can now prove the main theorem.

Theorem 5.5. *The semigroup $\mathcal{S}:[0,+\infty)\otimes L^\infty(\mathbb{R};E)\longmapsto L^\infty(\mathbb{R};E)$ defined in* (5.10) *is the only continuous semigroup on $L^\infty(\mathbb{R};E)$ such that the following properties are satisfied:*

(i) *for all* $\bar{u}^n, \bar{u} \in L^\infty(\mathbb{R}; E)$, $t_n, t \in [0, +\infty)$, *with* $\bar{u}_n \to \bar{u}$ *in* L^1_{loc}, $|t - t_n| \to 0$ *as* $n \to +\infty$,

$$\lim_{n \to +\infty} \mathcal{S}_{t_n} \bar{u}_n = \mathcal{S}_t \bar{u} \quad \textit{in } L^1_{\mathrm{loc}}; \tag{5.12}$$

(ii) *each trajectory* $t \mapsto \mathcal{S}_t u_0$ *is a weak entropic solution to the Cauchy problem*

$$\begin{cases} u_t + f(u)_x = 0, \\ u(0, x) = u_0(x) \end{cases} \tag{5.13}$$

with $u_0 \in L^\infty(\mathbb{R}; E)$;

(iii) *if* u_0 *is piecewise constant, then, for* t *sufficiently small,* $\mathcal{S}_t u_0$ *coincides with the function obtained by piecing together the solutions of the corresponding Riemann problems.*

Proof. The statement follows easily, since we proved that $\mathcal{S}_t u$ is the unique limit of wave front approximations, and for data with bounded total variation we can apply the results in [3]. □

Remark 5.6. *Note that we also proved that the characteristic equation* (4.1) *is well posed for* L^∞ *data: the solution* $x_i(t, y)$ *is Lipschitz continuous w.r.t. both variables. In fact, if* u_0^n *converges to* u_0 *in* L^1_{loc}, *Proposition* 4.2 *implies that* $x_i^n(t, y)$, *solution to the* i *characteristic equation, tends to* $x(t, y)$ *uniformly for all* t, y. *It is then easy to prove that* $x(t, y)$ *satisfies the equation*

$$x_i(t, y) = y + \int_0^t \lambda_i\big(s, x_i(s, y)\big) ds.$$

The above equation implies uniqueness of $x_i(t, y)$ *in the sense of Carathéodory, and Proposition* 4.3 *prove continuous dependence on* y.

This is not trivial, since even for 2×2 *systems not in conservation form the dependence is Hölder continuous, while for general* $n \times n$ *the solution does not exist* [16].

Note, moreover, that semigroup $\mathcal{S}$ *is continuous but not uniformly continuous. However, if the initial data takes values in a compact set of* $L^1 \cap L^\infty$, *then the semigroup becomes uniformly continuous. This extends the Lipschitz continuity when the initial data have bounded total variation.*

Acknowledgment. I wish to thank Prof. Alberto Bressan for having suggested the problem and his careful reading of the manuscript.

REFERENCES

[1] F. Ancona and A. Marson, *On the attainable set for scalar nonlinear conservation laws with boundary control*, SIAM J. Control Optim., 36 (1998), pp. 290–312.

[2] A. Aw and M. Rascle, *Resurrection of "second order" models of traffic flow*, SIAM J. Appl. Math., 60 (2000), pp. 916–938.

[3] P. Baiti and A. Bressan, *The semigroup generated by a Temple class system with large data*, Differential Integral Equations, 10 (1997), pp. 401–418.

[4] 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.

[5] S. Bianchini, *T*he semigroup generated by a Temple class system with non-convex flux function, Differential Integral Equations, 13 (2000), pp. 1529–1550.

[6] A. Bressan, *Hyperbolic Systems of Conservation Laws. The One Dimensional Cauchy Problem*, Oxford University Press, London, 2000.

[7] A. Bressan, *The unique limit of the Glimm scheme*, Arch. Ration. Mech. Anal., 130 (1995), pp. 205–230.

[8] A. Bressan, *Unique solutions for a class of discontinuous differential equations*, Proc. Amer. Math. Soc., 104 (1988), pp. 772–778.

[9] A. Bressan, G. Crasta, and B. Piccoli, *Well posedness of the Cauchy problem for $n \times n$ systems of conservation laws*, Mem. Amer. Math. Soc., 146 (2000).

[10] A. Bressan and R.M. Colombo, *The semigroup generated by 2×2 conservation laws*, Arch. Ration. Mech. Anal., 133 (1995), pp. 1–75.

[11] A. Bressan and P. Goatin, *Oleinik type estimates and uniqueness for $n \times n$ conservation laws*, J. Differential Equations, 156 (1999), pp. 26–49.

[12] A. Bressan and P. Goatin, *Stability of L^∞ solutions of Temple class systems*, Differential Integral Equations, 13 (2000), pp. 1503–1528.

[13] A. Bressan and P. LeFloch, *Uniqueness of weak solutions to hyperbolic systems of conservation laws*, Arch. Ration. Mech. Anal., 140 (1997), pp. 301–317.

[14] A. Bressan and M. Lewicka, *A uniqueness condition for hyperbolic systems of conservation laws*, Discrete Contin. Dynam. Systems, 6 (2000), pp. 673–682.

[15] A. Bressan, T.P. Liu, and T. Yang, *L^1 stability estimates for $n \times n$ conservation laws*, Arch. Ration. Mech. Anal., 149 (1999), pp. 1–22.

[16] A. Bressan and W. Shen, *Uniqueness for discontinuous O.D.E. and conservation laws*, Nonlinear Anal., 34 (1998), pp. 637–652.

[17] R.J. DiPerna, *Convergence of approximate solutions to conservation laws*, Arch. Ration. Mech. Anal., 82 (1983), pp. 27–70.

[18] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.

[19] J. Glimm and P. Lax, *Decay of solutions of systems of nonlinear hyperbolic conservation laws*, Mem. Amer. Math. Soc., 101 (1970).

[20] H.K. Jenssen, *Blowup for systems of conservation laws*, SIAM J. Math. Anal., 31 (2000), pp. 894–908.

[21] S. Kruzhkov, *First-order quasilinear equations with several space variables*, Mat. Sb., 123 (1970), pp. 228–255 (in Russian); Math. USSR Sb., 10 (1970), pp. 217–273 (in English).

[22] P. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[23] M. Lewicka, *On the well posedness of a system of balance laws with L^∞ data*, Rend. Sem. Mat. Univ. Padova, 102 (1999), pp. 319–340.

[24] D. Serre, *Solutions à variation bornée pour certains systèmes hyperboliques de lois de conservation*, J. Differential Equations, 67 (1983), pp. 137–168.

[25] D. Serre, *Systèmes de lois de conservation*, Diderot Editeur, Paris, 1996.

[26] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

[27] B. Temple, *Systems of conservation laws with invariant submanifolds*, Trans. Amer. Math. Soc., 280 (1983), pp. 781–795.

SIAM J. MATH. ANAL.
Vol. 33, No. 4, pp. 982–988

ILL-POSEDNESS ISSUES FOR THE BENJAMIN–ONO AND RELATED EQUATIONS[*]

L. MOLINET[†], J. C. SAUT[‡], AND N. TZVETKOV[‡]

Abstract. We establish that the Cauchy problem for the Benjamin–Ono equation and for a rather general class of nonlinear dispersive equations with dispersion slightly weaker than that of the Korteweg–de Vries equation cannot be solved by an iteration scheme based on the Duhamel formula. As a consequence, the flow map fails to be smooth.

Key words. Benjamin–Ono equation, internal long waves, Cauchy problem, ill-posedness

AMS subject classifications. 35Q53, 35B30, 76B15, 76C10

PII. S0036141001385307

1. Introduction.

The Benjamin–Ono equation is one of the fundamental equations describing the evolution of weakly nonlinear dispersive internal long waves. It has been derived by Benjamin [4] as an approximate model for long-crested unidirectional waves at the interface of a two-layer system of incompressible inviscid fluids; the top layer is assumed to be infinitely deep, while the heavier bottom layer has finite depth. In nondimensional variables, the Benjamin–Ono equation writes as

$$u_t - Hu_{xx} + uu_x = 0, \quad u = u(t,x), \quad (t,x) \in \mathbb{R}^2, \tag{1}$$

where $u(t,x)$ is proportional to the vertical deviation of the interface from its rest position at the point x at time t, and H is the Hilbert transform applied in the spatial variable.

The Benjamin–Ono equation has attracted a considerable number of papers during the last 20 years, particularly because it is completely integrable, at least formally [3]. Actually, rigorous results concerning the inverse scattering transform method require a certain small norm condition that excludes solitary waves [8].

We now review the state of the art concerning the Cauchy problem for the Benjamin–Ono equation, which turns out to be more delicate than expected.

First, (1) possesses two conservation laws:[1]

$$\int_{-\infty}^{\infty} u^2(t,x)dx = \int_{-\infty}^{\infty} u^2(0,x)dx \qquad \text{(momentum)}, \tag{2}$$

$$\begin{aligned}\int_{-\infty}^{\infty} \left[\frac{1}{2}||D_x|^{\frac{1}{2}}u|^2 - \frac{1}{6}u^3\right](t,x)dx \\ = \int_{-\infty}^{\infty} \left[\frac{1}{2}||D_x|^{\frac{1}{2}}u|^2 - \frac{1}{6}u^3\right](0,x)dx \qquad \text{(energy)}.\end{aligned} \tag{3}$$

[*] Received by the editors February 20, 2001; accepted for publication (in revised form) June 5, 2001; published electronically December 28, 2001.
http://www.siam.org/journals/sima/33-4/38530.html

[†] L.A.G.A., Institut Galilée, Université Paris-Nord, 93430 Villetaneuse, France (molinet@math.univ-paris13.fr).

[‡] Université de Paris-Sud, UMR de Mathématiques, Bât. 425, 91405 Orsay Cedex, France.

[1] In fact, (1) possesses an infinite number due to the integrability.

This suggests that the space $L^2(\mathbb{R})$ or the energy space $H^{1/2}(\mathbb{R})$ are good candidates for a global well-posedness theory of the Cauchy problem. This problem is open and one aim of this paper is to present some obstructions to its solvability by iteration methods. On the other hand, some positive results are known. The Benjamin–Ono equation has global weak solutions in the energy space (see [19]) and even global weak solutions for data in L^2 (see [10, 11, 22]). This last result is based on smoothing properties of the underlying linear group. It has also been shown that the Cauchy problem is globally well-posed in H^s, $s > 3/2$ (see [9, 2]).[2] This result has been extended to $s = 3/2$ by Ponce (see [18]) by using dispersive estimates on the linear group. It is worth noticing that all the results summarized above are based on a compactness method: one passes to the limit by compactness arguments on some approximate solutions. Another natural way of solving the Cauchy problem is to perform an iterative method on the integral equation corresponding to the Benjamin–Ono equation with initial data ϕ:

$$u(t) = S(t)\phi - \int_0^t S(t-t')(u_x(t')u(t'))dt',$$

where $S(t) = \exp(tH\partial_x^2)$ is the generator of the free evolution.

Our results are negative ones and are inspired by a previous work of the authors on the Kadomtsev-Petviashvili-I equation (see [17]). They show that one cannot solve the Cauchy problem for the Benjamin–Ono equation by a Picard iterative method implemented on the integral formulation of (1) for initial data in the Sobolev space $H^s(\mathbb{R})$, $s \in \mathbb{R}$. In particular the methods introduced by Bourgain [6] and Kenig, Ponce, and Vega [13] for the Korteweg–de Vries (KdV) equation cannot be used for the Benjamin–Ono equation. Note that scaling arguments imply that the Cauchy problem should be ill-posed in $H^s(\mathbb{R})$, $s < -1/2$, and actually this has been recently proved in [5].

As a consequence, there does not exist a $T > 0$ such that (1) admits a unique local solution defined on the interval $[-T, T]$ and such that the flow-map data-solution $\phi \longmapsto u(t),\ t \in [-T, T]$, is C^2 differentiable at the origin from $H^s(\mathbb{R})$ to $H^s(\mathbb{R})$. This implies in particular that the flow map for the Cauchy problem solved in [2, 9, 18] is not smooth (C^2) at the origin, a fact which would be difficult to establish from the compactness theory alone.

We note that the situation is radically different for the generalized Benjamin–Ono equation

$$u_t + u^k u_x - Hu_{xx} = 0, \tag{4}$$

which has been studied by Kenig, Ponce, and Vega [14]. Actually, they establish by an iterative method (thus the corresponding flow is smooth) that (4) is locally well-posed for small data in a Sobolev space $H^{s_k}(\mathbb{R})$ ($s_k > 1$ if $k = 2$, $s_k > 5/6$ if $k = 3$, $s_k \geq 3/4$ if $k \geq 4$).

The rest of the paper is organized as follows. In the next section we state and prove our main results for the Benjamin–Ono equation. Section 3 is devoted to a natural extension of our arguments to a general class of nonlinear dispersive equations with dispersion slightly weaker than that of the KdV equation. This includes, for instance, the intermediate long wave equation (ILW), the Smith equation, and equations describing waves in rotating fluids.

[2] This result uses the next conservation law.

Notations. Different numerical constants are denoted by c and may change from line to line. For any positive A and B the notation $A \lesssim B$ (resp., $A \gtrsim B$) means that there exists a positive constant c such that $A \leq cB$ (resp., $A \geq cB$). The notation $A \sim B$ means that $A \lesssim B \lesssim A$. We denote the Fourier transform by $\widehat{\cdot}$ or $\mathcal{F}$.

2. The Benjamin–Ono equation. Consider the Cauchy problem

$$\left\{\begin{array}{l} u_t - Hu_{xx} + uu_x = 0,\ (t,x) \in \mathbb{R}^2, \\ u(0,x) = \phi(x). \end{array}\right. \tag{5}$$

Write (5) as an integral equation:

$$u(t) = S(t)\phi - \int_0^t S(t-t')(u_x(t')u(t'))dt'. \tag{6}$$

Our main results follow.

THEOREM 1. *Let $s \in \mathbb{R}$ and T be a positive real number. Then there does not exist a space X_T continuously embedded in $C([-T,T], H^s(\mathbb{R}))$ such that there exists $C > 0$ with*

$$\|S(t)\phi\|_{X_T} \leq C\|\phi\|_{H^s(\mathbb{R})}, \quad \phi \in H^s(\mathbb{R}), \tag{7}$$

and

$$\left\| \int_0^t S(t-t')\,[u(t')u_x(t')]\,dt' \right\|_{X_T} \leq C\|u\|^2_{X_T}, \quad u \in X_T. \tag{8}$$

Note that (7) and (8) would be needed to implement a Picard iterative scheme on (6), in the space X_T. As a consequence of Theorem 1 we can obtain the following result.

THEOREM 2. *Fix $s \in \mathbb{R}$. Then there does not exist a $T > 0$ such that* (5) *admits a unique local solution defined on the interval $[-T,T]$ and such that the flow-map data-solution*

$$\phi \longmapsto u(t), \quad t \in [-T,T],$$

for (5) *is C^2 differentiable at zero from $H^s(\mathbb{R})$ to $H^s(\mathbb{R})$.*

We note that in [23] the failure of C^2 regularity of the flow map of the KdV equation on $H^s(\mathbb{R})$ for $s < -3/4$ is studied, motivated by work of Bourgain [7]. A direct corollary of Theorem 2 is the next statement.

THEOREM 3. *The flow map in the existing results for the Benjamin–Ono equation is not C^2 from $H^s(\mathbb{R})$ to $H^s(\mathbb{R})$.*

2.1. Proof of Theorem 1. Suppose that there exists a space X_T such that (7) and (8) hold. Take $u = S(t)\phi$ in (8). Then

$$\left\| \int_0^t S(t-t')\,[(S(t')\phi)(S(t')\phi_x)]\,dt' \right\|_{X_T} \leq C\|S(t)\phi\|^2_{X_T}.$$

Now using (7) and that X_T is continuously embedded in $C([-T,T], H^s(\mathbb{R}))$ we obtain for any $t \in [-T,T]$ that

$$\left\| \int_0^t S(t-t')\,[(S(t')\phi)(S(t')\phi_x)]\,dt' \right\|_{H^s(\mathbb{R})} \lesssim \|\phi\|^2_{H^s(\mathbb{R})}. \tag{9}$$

We show that (9) fails by choosing an appropriate ϕ.

Take ϕ defined by its Fourier transform as[3]

$$\widehat{\phi}(\xi) = \alpha^{-\frac{1}{2}} \mathbb{1}_{I_1}(\xi) + \alpha^{-\frac{1}{2}} N^{-s} \mathbb{1}_{I_2}(\xi), \quad N \gg 1, \quad 0 < \alpha \ll 1,$$

where I_1, I_2 are the intervals

$$I_1 = [\alpha/2, \alpha], \quad I_2 = [N, N+\alpha].$$

Note that $\|\phi\|_{H^s} \sim 1$. We will use the next lemma.

LEMMA 1. *The following identity holds:*

$$\int_0^t S(t-t')\Big[(S(t')\phi)(S(t')\phi_x)\Big]\,dt'$$
$$= c\int_{\mathbb{R}^2} e^{ix\xi+itp(\xi)}\,\xi\,\hat{\phi}(\xi_1)\hat{\phi}(\xi-\xi_1)\frac{e^{it(p(\xi_1)+p(\xi-\xi_1)-p(\xi))}-1}{p(\xi_1)+p(\xi-\xi_1)-p(\xi)}d\xi d\xi_1,$$

where $p(\xi) = \xi|\xi|$.

Proof of Lemma 1. The proof is very similar to that of Lemma 4 in [17]. We give it for the sake of completeness. Taking the inverse Fourier transform with respect to x, it is easily seen that

$$\int_0^t S(t-t')\Big[(S(t')\phi)(S(t')\phi_x)\Big]\,dt'$$
$$= c\int_0^t\int_{\mathbb{R}} e^{ix\xi+itp(\xi)}e^{-it'p(\xi)}\xi\Big[(e^{it'p(\cdot)}\hat{\phi}(\cdot)) * (e^{it'p(\cdot)}\hat{\phi}(\cdot))\Big](\xi)\,d\xi\,dt'$$
$$= c\int_{\mathbb{R}^2} e^{ix\xi+itp(\xi)}\,\xi\,\hat{\phi}(\xi_1)\hat{\phi}(\xi-\xi_1)\int_0^t e^{it'(p(\xi_1)+p(\xi-\xi_1)-p(\xi))}\,dt'\,d\xi_1 d\xi$$
$$= c\int_{\mathbb{R}^2} e^{ix\xi+itp(\xi)}\,\xi\,\hat{\phi}(\xi_1)\hat{\phi}(\xi-\xi_1)\frac{e^{it(p(\xi_1)+p(\xi-\xi_1)-p(\xi))}-1}{p(\xi_1)+p(\xi-\xi_1)-p(\xi)}\,d\xi_1 d\xi. \qquad \square$$

According to the above lemma,

$$\int_0^t S(t-t')\,[(S(t')\phi)(S(t')\phi_x)]\,dt' = c(f_1(t,x)+f_2(t,x)+f_3(t,x)),$$

where, from the definition of ϕ, we have the following representations for f_1, f_2, f_3:

$$f_1(t,x) = \frac{c}{\alpha}\int_{\substack{\xi_1\in I_1\\ \xi-\xi_1\in I_1}} \xi\, e^{ix\xi+it\xi|\xi|}\frac{e^{it(\xi_1|\xi_1|+(\xi-\xi_1)|\xi-\xi_1|-\xi|\xi|)}-1}{\xi_1|\xi_1|+(\xi-\xi_1)|\xi-\xi_1|-\xi|\xi|}d\xi d\xi_1,$$

$$f_2(t,x) = \frac{c}{\alpha N^{2s}}\int_{\substack{\xi_1\in I_2\\ \xi-\xi_1\in I_2}} \xi\, e^{ix\xi+it\xi|\xi|}\frac{e^{it(\xi_1|\xi_1|+(\xi-\xi_1)|\xi-\xi_1|-\xi|\xi|)}-1}{\xi_1|\xi_1|+(\xi-\xi_1)|\xi-\xi_1|-\xi|\xi|}d\xi d\xi_1,$$

$$f_3(t,x) = \frac{c}{\alpha N^{s}}\int_{\substack{\xi_1\in I_1\\ \xi-\xi_1\in I_2}} \xi\, e^{ix\xi+it\xi|\xi|}\frac{e^{it(\xi_1|\xi_1|+(\xi-\xi_1)|\xi-\xi_1|-\xi|\xi|)}-1}{\xi_1|\xi_1|+(\xi-\xi_1)|\xi-\xi_1|-\xi|\xi|}d\xi d\xi_1$$

$$+ \quad \frac{c}{\alpha N^{s}}\int_{\substack{\xi_1\in I_2\\ \xi-\xi_1\in I_1}} \xi\, e^{ix\xi+it\xi|\xi|}\frac{e^{it(\xi_1|\xi_1|+(\xi-\xi_1)|\xi-\xi_1|-\xi|\xi|)}-1}{\xi_1|\xi_1|+(\xi-\xi_1)|\xi-\xi_1|-\xi|\xi|}d\xi d\xi_1.$$

[3] The analysis below works as well for $\mathcal{R}e\,\phi$ instead of ϕ (some new harmless terms appear).

Set

$$\chi(\xi,\xi_1) := \xi_1|\xi_1| + (\xi-\xi_1)|\xi-\xi_1| - \xi|\xi|.$$

Then clearly

$$\mathcal{F}_{x\mapsto\xi}(f_1)(t,\xi) = \frac{c\,\xi e^{it\xi|\xi|}}{\alpha}\int_{\substack{\xi_1\in I_1\\ \xi-\xi_1\in I_1}} \frac{e^{it\chi(\xi,\xi_1)}-1}{\chi(\xi,\xi_1)}d\xi_1,$$

$$\mathcal{F}_{x\mapsto\xi}(f_2)(t,\xi) = \frac{c\,\xi e^{it\xi|\xi|}}{\alpha N^{2s}}\int_{\substack{\xi_1\in I_2\\ \xi-\xi_1\in I_2}} \frac{e^{it\chi(\xi,\xi_1)}-1}{\chi(\xi,\xi_1)}d\xi_1,$$

$$\mathcal{F}_{x\mapsto\xi}(f_3)(t,\xi) = \frac{c\,\xi e^{it\xi|\xi|}}{\alpha N^{s}}\left(\int_{\substack{\xi_1\in I_1\\ \xi-\xi_1\in I_2}} \frac{e^{it\chi(\xi,\xi_1)}-1}{\chi(\xi,\xi_1)}d\xi_1 + \int_{\substack{\xi_1\in I_2\\ \xi-\xi_1\in I_1}} \frac{e^{it\chi(\xi,\xi_1)}-1}{\chi(\xi,\xi_1)}d\xi_1\right).$$

Since the supports of $\mathcal{F}_{x\mapsto\xi}(f_j)(t,\xi)$, $j=1,2,3$, are disjoint, we have

$$\left\|\int_0^t S(t-t')\left[(S(t')\phi)(S(t')\phi_x)\right]dt'\right\|_{H^s(\mathbb{R})} \geq \|f_3(t,\cdot)\|_{H^s(\mathbb{R})}.$$

We now give a lower bound for $\|f_3(t,\cdot)\|_{H^s(\mathbb{R})}$. Note that for $(\xi_1,\xi-\xi_1)\in I_1\times I_2$ or $(\xi_1,\xi-\xi_1)\in I_2\times I_1$ one has $|\chi(\xi,\xi_1)| = 2|\xi_1(\xi-\xi_1)| \sim \alpha N$. Hence it is natural to choose α and N so that $\alpha N = N^{-\epsilon}$, $0<\epsilon\ll 1$. Then

$$\left|\frac{e^{it\chi(\xi,\xi_1)}-1}{\chi(\xi,\xi_1)}\right| = |t| + O(N^{-\epsilon})$$

for $\xi_1\in I_1$, $\xi-\xi_1\in I_2$ or $\xi_1\in I_2$, $\xi-\xi_1\in I_1$. Hence for $t\neq 0$,

$$\|f_3(t,\cdot)\|_{H^s(\mathbb{R})} \gtrsim \frac{N\,N^s\,\alpha\,\alpha^{\frac{1}{2}}}{\alpha N^s} = \alpha^{\frac{1}{2}}N.$$

Therefore we arrive at

$$1\sim\|\phi\|^2_{H^s(\mathbb{R})} \geq \|f_3(t,\cdot)\|_{H^s(\mathbb{R})} \geq \alpha^{\frac{1}{2}}N \sim N^{\frac{1-\epsilon}{2}},$$

which is a contradiction for $N\gg 1$ and $\epsilon\ll 1$. This completes the proof of Theorem 1. □

2.2. Proof of Theorem 2. Consider the Cauchy problem

$$\begin{cases} u_t - Hu_{xx} + uu_x = 0, \\ u(0,x) = \gamma\phi, \quad \gamma\ll 1, \quad \phi\in H^s(\mathbb{R}). \end{cases} \tag{10}$$

Suppose that $u(\gamma,t,x)$ is a local solution of (10) and that the flow map is C^2 at the origin from $H^s(\mathbb{R})$ to $H^s(\mathbb{R})$. We have

$$\frac{\partial^2 u}{\partial\gamma^2}(0,t,x) = -2\int_0^t S(t-t')\left[(S(t')\phi)(S(t')\phi_x)\right]dt'.$$

The assumption of C^2 regularity yields

$$\left\|\int_0^t S(t-t')\left[(S(t')\phi)(S(t')\phi_x)\right]dt'\right\|_{H^s(\mathbb{R})} \lesssim \|\phi\|^2_{H^s(\mathbb{R})}.$$

But the above estimate is (9), which has been shown to fail in section 2.1. □

3. A class of nonlinear dispersive equations. We consider now the class of equations

$$u_t + uu_x - Lu_x = 0, \quad u(0,x) = \phi(x), \quad (t,x) \in \mathbb{R}^2, \tag{11}$$

where L is defined via the Fourier transform

$$\widehat{Lf}(\xi) = \omega(\xi)\hat{f}(\xi).$$

Here $\omega(\xi)$ is a continuous real-valued function. Set $p(\xi) = \xi\,\omega(\xi)$. We assume that $p(\xi)$ is differentiable and such that, for some $\gamma \in \mathbb{R}$,

$$|p'(\xi)| \lesssim |\xi|^\gamma, \quad \xi \in \mathbb{R}. \tag{12}$$

The next theorem shows that (11) shares the bad behavior of the Benjamin–Ono equation with respect to iterative methods.

Theorem 4. *Assume that* (12) *holds with* $\gamma \in [0, 2[$. *Then the conclusions of Theorems* 1, 2, *and* 3 *are valid for the Cauchy problem* (11).

The proof follows the considerations of the previous section. The main point in the analysis is that for $\xi_1 \in I_1$, $\xi - \xi_1 \in I_2$ one has

$$|p(\xi_1) + p(\xi - \xi_1) - p(\xi)| \lesssim \alpha N^\gamma, \qquad \alpha \ll 1, \quad N \gg 1.$$

We choose α and N such that $\alpha N^\gamma = N^{-\epsilon}$, $0 < \epsilon \ll 1$. We take the same ϕ as in the proof of Theorem 1 and arrive at the lower bound

$$1 \sim \|\phi\|^2_{H^s(\mathbb{R})} \geq \alpha^{\frac{1}{2}} N = N^{1-\frac{\gamma+\epsilon}{2}},$$

which fails for $0 < \epsilon \ll 1$, $\gamma \in [0, 2[$.

Here we give several examples where Theorem 4 applies.

- Pure power dispersion:

$$\omega(\xi) = |\xi|^\gamma, \quad 0 \leq \gamma < 2.$$

This dispersion corresponds to a class of models for vorticity waves in the coastal zone (see [20]). It is interesting to notice that the case $\gamma = 2$ corresponds to the KdV equation which can be solved by iterative methods (see [6, 13]). Therefore Theorem 4 is sharp for a pure power dispersion. However, the Cauchy problem corresponding to $1 \leq \gamma < 2$ has been proven in [12, Theorem 1.3] to be locally well-posed by a compactness method combined with sharp estimates on the linear group for initial data in $H^s(\mathbb{R})$, $s \geq (9 - 3\gamma)/4$.

- Perturbations of the Benjamin–Ono equation:

$$\omega(\xi) = (|\xi|^2 + 1)^{\frac{1}{2}} \quad \text{(see [21])},$$

$$\omega(\xi) = \xi \coth(\xi) \quad \text{(the ILW equation; see [15, 1, 2])}.$$

- Equations describing waves in rotating fluids:

$$\omega(\xi) = \xi^2 K_0(|\xi|) \quad \text{(where } K_0 \text{ is the Bessel function of order zero; see [16])}.$$

REFERENCES

[1] J. Albert, J. Bona, and J.C. Saut, *Model equations for waves in stratified fluids*, Proc. Roy. Soc. London Ser. A, 453 (1997), pp. 1233–1260.

[2] L. Abdelouhab, J. Bona, M. Felland, and J.C. Saut, *Nonlocal models for nonlinear, dispersive waves*, Phys. D, 40 (1989), pp. 360–392.

[3] M.J. Ablowitz and A.S. Fokas, *The inverse scattering transform for the Benjamin-Ono equation, a pivot for multidimensional problems*, Stud. Appl. Math., 68 (1983), pp. 1–10.

[4] T.B. Benjamin, *Internal waves of permanent form in fluids of great depth*, J. Fluid Mech., 29 (1967), pp. 559–592.

[5] H.A. Biagioni and F. Linares, *Ill-Posedness for the Derivative Schrödinger and Generalized Benjamin-Ono Equations*, preprint, 2000.

[6] J. Bourgain, *Fourier transform restriction phenomena for certain lattice subsets and application to nonlinear evolution equations* II. *The KdV equation*, Geom. Funct. Anal., 3 (1993), pp. 209–262.

[7] J. Bourgain, *Periodic Korteweg de Vries equation with measures as initial data*, Selecta Math. (N.S.), 3 (1997), pp. 115–159.

[8] R. Coifman and M. Wickerhauser, *The scattering transform for the Benjamin-Ono equation*, Inverse Problems, 6 (1990), pp. 825–860.

[9] R. Iorio, *On the Cauchy problem for the Benjamin-Ono equation*, Comm. Partial Differential Equations, 11 (1986), pp. 1031–1081.

[10] J. Ginibre and G. Velo, *Propriétés de lissage et existence de solutions pour l'équation de Benjamin-Ono généralisée*, C.R. Acad. Sci. Paris Sér. I Math., 308 (1989), pp. 309–314.

[11] J. Ginibre and G. Velo, *Smoothing properties and existence of solutions for the generalized Benjamin-Ono equation*, J. Differential Equations, 93 (1991), pp. 150–232.

[12] C. Kenig, G. Ponce, and L. Vega, *Well-posedness of the initial value problem for the Korteweg-de Vries equation*, J. Amer. Math. Soc., 4 (1991), pp. 323–347.

[13] C. Kenig, G. Ponce, and L. Vega, *Well-posedness and scattering results for the generalized KdV equations via the contraction principle*, Comm. Pure Appl. Math., 46 (1993), pp. 527–620.

[14] C. Kenig, G. Ponce, and L. Vega, *On the generalized Benjamin-Ono equation*, Trans. Amer. Math. Soc., 342 (1994), pp. 155–172.

[15] T. Kubota, D. Ko, and L. Dobbs, *Weakly nonlinear internal gravity waves in stratified fluids of finite depth*, J. Hydrodynamics, 12 (1978), pp. 157–165.

[16] S. Leibovitch, *Weakly nonlinear waves in rotating fluids*, J. Fluid Mech., 42 (1970), pp. 803–822.

[17] L. Molinet, J.C. Saut, and N. Tzvetkov, *Well-posedness and ill-posedness results for the Kadomtsev-Petviashvili-I equation*, Duke Math J., to appear.

[18] G. Ponce, *On the global well-posedness of the Benjamin-Ono equation*, Differential Integral Equations, 4 (1991), pp. 527–542.

[19] J.C. Saut, *Sur quelques généralisations de l'équation de Korteweg-de Vries*, J. Math. Pures Appl., 58 (1979), pp. 21–61.

[20] V.I. Shrira and V.I. Voronovich, *Nonlinear dynamics of vorticity waves in the coastal zone*, J. Fluid Mech., 326 (1996), pp. 181–203.

[21] R. Smith, *Nonlinear Kelvin and continental-shelf waves*, J. Fluid Mech., 57 (1972), pp. 379–391.

[22] M. Tom, *Smoothing properties of some weak solutions of the Benjamin-Ono equation*, Differential Integral Equations, 3 (1990), pp. 683–694.

[23] N. Tzvetkov, *Remark on the local ill-posedness for KdV equation*, C.R. Acad. Sci. Paris Sér. I Math., 329 (1999), pp. 1043–1047.

SIAM J. Math. Anal.
Vol. 33, No. 4, pp. 989–993

A LOCAL INVERSION PRINCIPLE OF THE NASH–MOSER TYPE*

ALFONSO CASTRO† AND J. W. NEUBERGER‡

Abstract. We prove an inverse function theorem of the Nash–Moser type. The main difference between our method and that of [J. Moser, *Proc. Nat. Acad. Sci. USA,* 47 (1961), pp. 1824–1831] is that we use continuous steepest descent while Moser uses a combination of Newton-type iterations and approximate inverses. We bypass the *loss of derivatives problem* by working on finite dimensional subspaces of infinitely differentiable functions.

Key words. Nash–Moser methods, inverse function theorem, continuous steepest descent

AMS subject classifications. Primary, 34B15; Secondary, 35J65

PII. S0036141099357690

1. Introduction.

Inverse function theorems are fundamental tools for the study of solutions to nonlinear equations. Proofs depend on iteration arguments. When a nonlinear equation comes from a partial differential equation, it often happens that the operators under consideration do not have enough *regularity* and the iteration process is defined only for a few steps. This is known as loss of derivatives. In order to overcome this difficulty, iteration techniques have been designed to allow the iteration to lead to a limit. These are known as generalized inverse function theorems. They generally depend on having a *scale* of intermediate spaces between the domain and the range of the nonlinear operator under consideration. Most notable of such results is the one proven by J. Moser in [5] and [6] (see also [2]). The reader is referred to [3] for a sharper version of the results of [5].

Here we prove an inverse function theorem (Theorem 2.1 below) using finite dimensional subspaces of infinitely differentiable functions, which makes the phenomenon of loss of derivatives immaterial. In addition we only assume the operators to have a first order derivative. This is in contrast with the results of [5] and [3] where Newton-like iteration techniques require extensive use of the properties of the *quadratic remainder* (see (10)). In [12] another inverse function theorem is proven without assumptions on the quadratic remainder. Our use of continuous steepest descent has roots in [1], [7], [8], and [9]. For applications of generalized inverse function theorems to elliptic systems the reader is referred to [4] and [11].

2. Main result.

For the sake of simplicity we present our main result in the context of a particular class of Sobolev spaces. However, the general principle applies to many other *scales* of spaces sharing the general properties of the Sobolev spaces that are defined in the next paragraph. The reader is referred to [2] for various other scales of spaces.

Let each of m and n denote a positive integer. For each nonnegative integer ρ let C^ρ denote the set of ρ-differentiable functions $u : R^n \to R^m$ that are 2π-periodic in each of its n independent real variables. The norm in C^ρ is given by

$$|u|_\rho = \max\{|u(x)|; x \in R^n\} + \max\{|D^\alpha u(x)|; x \in R^n, |\alpha| = \rho\}; \tag{1}$$

*Received by the editors June 23, 1999; accepted for publication (in revised form) June 7, 2001; published electronically December 28, 2001.

http://www.siam.org/journals/sima/33-4/35769.html

†Division of Mathematics and Statistics, University of Texas at San Antonio, San Antonio, TX 78249-0664 (acastro@utsa.edu).

‡Department of Mathematics, University of North Texas, Denton, TX 76203 (jwn@unt.edu).

here, $\alpha = (\alpha_1, \dots, \alpha_n)$ is an n-tuple of nonnegative integers and $|\alpha| = \alpha_1 + \cdots + \alpha_n$. Similarly for $r \geq 0$ we define H^r as the Sobolev space of functions $u : R^n \to R^m$ of the form

$$u(x) = \sum_{j \in \mathbf{Z}^n} c_j e^{ij \cdot x} \tag{2}$$

such that

$$\|u\|_r^2 \equiv \sum_{j \in \mathbf{Z}^n} |c_j|^2 \ + \sum_{j \in \mathbf{Z}^n} |j|^{2r} |c_j|^2 < \infty, \tag{3}$$

where $|(j_1, \dots, j_n)|^2 = j_1^2 + \cdots + j_n^2$. Actually H^r is also defined for $r < 0$ provided that the expression in (2) is understood in the sense of distributions and

$$\|u\|_r^2 \equiv |c_0|^2 \ + \sum_{j \in \mathbf{Z}^n - \{0\}} |j|^{2r} |c_j|^2 < \infty. \tag{4}$$

For $r \geq 0$ the inner product in H^r is given by

$$\left\langle \sum_{j \in \mathbf{Z}^n} c_j e^{ij \cdot x}, \sum_{j \in \mathbf{Z}^n} d_j e^{ij \cdot x} \right\rangle_r = \sum_{j \in \mathbf{Z}^n} (1 + |j|^{2r}) c_j \bar{d}_j. \tag{5}$$

Let $\epsilon > 0$ and $F : \{x \in C^1; |x|_1 < \epsilon\} \to C^0$ be a continuous function such that $F(0) = 0$. Typically F is a first order differential operator. We assume that for $u, v \in C^1$, with $|u|_1 < \epsilon$, the limit

$$\lim_{t \to 0} \frac{F(u + tv) - F(u)}{t} \equiv F'(u)v \tag{6}$$

exists and is a continuous function of v and that $F'(\cdot)v$ defines a continuous function from C^1 into C^0, for each $v \in C^1$. We also assume that there exist $\lambda \in R$, positive constants k_1, k_2, k_3, and $l > (n/2) + 1$ with

$$\langle F'(u)v, v \rangle_0 \geq k_1 \|v\|_0^2 \tag{7}$$

for all $u, v \in C^1$;

$$\langle F'(u)v, v \rangle_l \geq k_2 \|v\|_\lambda^2 - k_3 \|v\|_0^2 \tag{8}$$

for $u \in C^1$, $|u|_l \leq \epsilon$, $v \in H^l$; and for each positive integer ρ there exists M_ρ such that

$$\|F(u)\|_\rho \leq M_\rho \|u\|_{\rho+1} \quad \text{for} \quad |u|_1 < \epsilon. \tag{9}$$

Without loss of generality we may assume that $k_1 < k_2$.

In [6], it is shown that the equation $F(u) = g$ has a solution when the operators $F'(u)$ have *approximate inverses*, and

$$\|F(u + v) - F(u) - F'(u)v\|_0 \leq M \|v\|_0^{2-\beta} \|v\|_l^\beta \ \text{ for } \ |u|_1 < \epsilon, \ |u + v|_1 < \epsilon, \tag{10}$$

where M is a constant independent of u, v and $0 \leq \beta < 1$. In addition, it is shown that (7) and (8) with $l = \lambda$ are sufficient for $F'(u)$ to have an approximate inverse.

Here we prove the following theorem.

THEOREM 2.1. *If (6)–(9) hold and $l > (n/2) + 1$, then there exist $\delta > 0$ such that, if $\|g\|_l < \delta$, then the equation*

$$F(y) = g \tag{11}$$

has a solution.

Proof. For each positive integer k, let X_k denote the linear subspace of H^l of functions of the form

$$u(x) = \sum_{\|j\|^2 \leq k^2} c_j e^{ij\cdot x}. \tag{12}$$

Let $P_k \equiv P$ denote the orthogonal projection of H^l onto X_k. An elementary Fourier series argument shows that P is also an orthogonal projection of H^0 onto X_k and

$$\langle Pu, v\rangle_0 = \langle u, Pv\rangle_0 \ \text{ for all } \quad u, v \in H^l. \tag{13}$$

Now let $\Lambda \in (n/2+1, l)$. Since X_k is finite dimensional and a subset of C^1, by (6) there exists a bounded differentiable function $\tilde{F} : X_k \to X_k$ such that $PF(u) = \tilde{F}(u)$ if $\|u\|_\Lambda < \epsilon/2$, $u \in X_k$. Hence the initial value problem

$$z'(t) = -\tilde{F}(z(t)) + P(g),\ t \geq 0, \quad z(0) = 0, \tag{14}$$

has a solution defined on $[0, \infty)$.

Let us see that, for $\|g\|_l$ small enough, $|z(t)|_1 < \epsilon/2$ for all $t \geq 0$. In fact, let $w(t) = P(\tilde{F}(z(t)) - g)$. Thus

$$\begin{aligned}(\|w(t)\|_0^2)' &= 2\langle w(t), P\tilde{F}'(z(t))z'(t)\rangle_0 \\ &= -2\langle Pw(t), \tilde{F}'(z(t))(P(w(t))\rangle_0 \\ &\leq -2k_1\|w(t)\|_0^2.\end{aligned} \tag{15}$$

In particular we see that the quantity $\|w(t)\|_0$ is a decreasing function of t. In addition, from (15) we have

$$\|w(t)\|_0 \leq \|w(0)\|_0 e^{-k_1 t} = \|g\|_0 e^{-k_1 t}. \tag{16}$$

Now we estimate the H^l norm of $w(t)$. In order to do so we observe that for each (k, λ) there exists a positive constant $C(k, \lambda)$ such that

$$\|x\|_\lambda^2 \geq C(k, \lambda)\|x\|_l^2 \ \text{ for all } \quad x \in X_k. \tag{17}$$

We note that $C(k, \lambda) \to 0$ as $k \to \infty$ when $\lambda < l$; otherwise $C(k, \lambda)$ can be taken to be equal to 1. From now on we restrict ourselves to the case $\lambda < l$; the case $\lambda \geq l$ is simpler. Thus we may assume that

$$C(k, \lambda) \to 0 \ \text{ as } \quad k \to \infty. \tag{18}$$

From (8), (13), (16), (17), and (18) we infer that for k sufficiently large

$$\begin{aligned}(\|w(t)\|_l^2)' &= 2\langle w(t), P\tilde{F}'(z(t))z'(t)\rangle_l \\ &= -2\langle Pw(t), \tilde{F}'(z(t))(P(w(t))\rangle_l \\ &\leq -2(k_2\|w(t)\|_\lambda^2 - k_3\|w(t)\|_0^2) \\ &\leq -2(k_2 C(k, \lambda)\|w(t)\|_l^2 - k_3\|g\|_0^2 e^{-2k_1 t}).\end{aligned} \tag{19}$$

Thus for k sufficiently large

$$\begin{aligned} \|w(t)\|_l^2 &\le (\|g\|_l^2 + k_3\|g\|_0^2/(k_1 - k_2C(k,\lambda)))e^{-2k_2C(k,\lambda)t} \\ &\le (\|g\|_l^2 + 2k_3\|g\|_0^2/k_1)e^{-2k_2C(k,\lambda)t} \\ &\equiv M(\|g\|_l)e^{-2k_2C(k,\lambda)t}, \end{aligned} \tag{20}$$

where $M(\|g\|_l) \to 0$ as $\|g\|_l \to 0$. By interpolation properties of Sobolev space (see section I.2 in [6]) one has

$$\|w(t)\|_\Lambda \le \|w(t)\|_0^{(1-\Lambda/l)}\|w(t)\|_l^{\Lambda/l} \le M(\|g\|_l)^{\Lambda/(2l)}\|g\|_0^{(1-\Lambda/l)}e^{-k_1(1-\Lambda/l)t}. \tag{21}$$

Integrating (14) by (20) and (21), we see that there exist $\delta > 0$ such that if $\|g\|_l \le \delta$, then

$$\|z(t)\|_\Lambda < \epsilon/3 \quad \text{for all} \quad t \ge 0. \tag{22}$$

Now letting $x_k \in X_k$ be an element in the w-limit set of the orbit defined by z, we see that $\|x_k\|_\Lambda \le \epsilon/3$ and $\tilde{F}(x_k) = PF(x_k) = P(g)$. Therefore by the Sobolev imbedding theorem we may assume that it converges to some element $x \in H^{\lambda_1}$, with $\lambda_1 \in ((N/2)+1, l)$. By (9), $F(x_k)$ is bounded in $H^{\lambda_1 - 1}$. Using again that bounded sequences in H^s have converging subsequences in H^t if $s > t$, we may further assume that $\{F(x_k)\}$ converges in H^{λ_2} with $\lambda_2 \in (n/2, \lambda_1 - 1)$. Recalling that, by Poincaré's inequality,

$$\|z\|_{\lambda_2 - 1} \le C_k\|z\|_{\lambda_2} \quad \text{for all} \quad z \in X_k^{\perp}, \tag{23}$$

with C_k converging to zero as $k \to \infty$, we conclude that $\|(I-P)F(x_k)\|_{\lambda_2 - 1} \to 0$ as $k \to \infty$. This and the fact that $\{F(x_k)\}$ converges to $F(x)$ imply that $\{P(F(x_k))\}$ converges to $F(x)$ in H^{λ_2}. Hence in H^{λ_2} we have

$$g = \lim P_k(g) = \lim P_k(F(x_k)) = F(x), \tag{24}$$

and this proves the theorem. □

REFERENCES

[1] A. Castro and J. W. Neuberger, *An inverse function theorem,* Contemp. Math., 221 (1999), pp. 127–132.

[2] R. Hamilton, *The inverse function theorem of Nash and Moser*, Bull. Amer. Math. Soc. (N.S.), 7 (1982), pp. 65–222.

[3] L. Hörmander, *On the Nash-Moser implicit function theorem*, Ann. Acad. Sci. Fenn. Ser. A. I. Math., 10 (1985), pp. 255–259.

[4] P. Korman, *On existence of solutions to two classes of nonlinear problems*, Comm. Partial Differential Equations, 14 (1989), pp. 519–539.

[5] J. Moser, *A new technique for the construction of solutions of nonlinear differential equations*, Proc. Nat. Acad. Sci., USA, 47 (1961), pp. 1824–1831.

[6] J. Moser, *A rapidly convergent iteration method and non-linear differential equations. I.*, Ann. Scuola Norm. Sup. Pisa (3), 20 (1966), pp. 265–315.

[7] J. W. Neuberger, *Steepest descent and differential equations*, J. Math. Soc. Japan, 37 (1985), pp. 187–195.

[8] J. W. Neuberger, *Constructive variational methods for differential equations*, Nonlinear Anal., 13 (1989), pp. 413–428.

[9] J. W. Neuberger, *Sobolev Gradients and Differential Equations*, Lecture Notes in Math. 1670, Springer-Verlag, Berlin, 1997.

[10] M. Poppenberg, *An inverse function theorem for Fréchet spaces satisfying smoothing property and (DN)*, Math. Nachr., 206 (1999), pp. 123–145.

[11] P. Rabinowitz, *A rapid convergence method for a singular perturbation problem*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 1 (1984), pp. 1–17.

[12] W. O. Ray, *A rapidly convergent iteration method and Gâteaux differentiable operators*, J. Math. Appl., 103 (1984), pp. 162–171.

SIAM J. Math. Anal.
Vol. 33, No. 5, pp. 995–1006

ON THE CONCEPT OF VERY WEAK L^2 SOLUTIONS TO EULER'S EQUATIONS[*]

HAMID BELLOUT[†], EMIL CORNEA[†], AND JINDŘICH NEČAS[‡]

Abstract. We propose a new concept of very weak solutions to Euler's equations. We show global existence of such solutions.

Key words. incompressible, fluid, Euler equations

AMS subject classifications. Primary, 35Q30, 35Q35; Secondary, 76B03

PII. S0036141000384290

1. Introduction.

The question of existence of global L^2 weak solutions in $\mathbf{R}^3$ to the initial boundary value problem for inviscid, incompressible fluids, governed by Euler equations

$$\rho \frac{\partial v_i}{\partial t} + \rho v_j \frac{\partial v_i}{\partial x_j} + \frac{\partial p}{\partial x_i} = 0 \quad \text{in } (0,T)\times\Omega, \quad i = 1,2,3, \tag{1.1}$$

$$\operatorname{div} \boldsymbol{v} = 0 \quad \text{in } (0,T)\times\Omega, \tag{1.2}$$

$$\boldsymbol{v}\cdot\boldsymbol{\nu} = 0 \quad \text{in } (0,T)\times\partial\Omega, \tag{1.3}$$

with initial condition

$$\boldsymbol{v}(0,\boldsymbol{x}) = \boldsymbol{v}^0(\boldsymbol{x}), \qquad \boldsymbol{x}\in\Omega, \tag{1.4}$$

where ρ is a positive constant (which will be assumed to be 1), is still open. A very nice and informative discussion of this question and related issues can be found in [4].

A concept of measure-valued solutions was proposed by DiPerna, Majda, Tartar, and others. See [4] and [5] and the references therein. This latter concept seems to be much too large a class of solutions. To quote Lions in [4, p. 153], *A very weak notion (relying on relaxed Young measures or relaxed measure valued solutions) is proposed by R.J. DiPerna and A. Majda* [129] *but the relevance of this notion is not entirely clear since it is not known that "solutions" in the sense of* [129] *coincide with smooth solutions as long as the latter do exist.* Another concept of solutions, called dissipative solutions, was introduced by Lions in [4].

For other results on this problem see also [1], [6], and the references therein.

We shall present here a straightforward concept, based on the *fixed point* for the map $\gamma : \boldsymbol{u} \to \boldsymbol{v}$, where $\boldsymbol{u}$ and $\gamma(\boldsymbol{u}) = \boldsymbol{v}$ are related by the Oseen equations

$$\frac{\partial v_i}{\partial t} + u_j \frac{\partial v_i}{\partial x_j} + \frac{\partial p}{\partial x_i} = 0 \quad \text{in } (0,T)\times\Omega, \quad i = 1,2,3, \tag{1.5}$$

[*] Received by the editors September 15, 2000; accepted for publication (in revised form) May 17, 2001; published electronically January 16, 2002. This research was supported, in part, by ONR grant N00014-95-1-0438 and by DoD grant N00014-97-1-0806.

http://www.siam.org/journals/sima/33-5/38429.html

[†] Department of Mathematical Sciences, Northern Illinois University, DeKalb, IL 60115 (bellout@math.niu.edu, cornea@math.niu.edu).

[‡] Mathematical Institute, Charles University, Prague, Czech Republic, and Department of Mathematical Sciences, Northern Illinois University, DeKalb, IL 60115 (necas@math.niu.edu).

where the fixed point is understood in the sense of the closure of the graph $(\boldsymbol{u}, \boldsymbol{v})$ in the topology of $L^{\infty}_{weak-star}((0,T); L^2_{weak}(\Omega)^3)$. Obviously any L^2 weak solution is a very weak solution in our sense.

We would like to make a few comments on the relation between the very weak solutions we introduce and other concepts of solutions.

1. It is not excluded that nonsmooth, very weak solutions coexist with classical solutions.
2. If a smooth classical solution exists on a time interval $[0, T_c)$, then the very weak solution provided by our method of construction will coincide with this classical solution over the time interval $[0, T_c)$.
3. If a very weak solution is known to be smooth in a time interval $[0, T_w)$, it will coincide with the classical solution over the time interval $[0, T_w)$.
4. Our very weak solutions and the dissipative solutions of Lions are in L^2, a space strictly smaller than the set where Young-measure solutions, whose existence was proved by Diperna and Majda, live.

We will prove items 2 and 3. In the proof of item 3 we will show that if an element $(\boldsymbol{u}, \boldsymbol{u})$ is in the closure of the graph for the topology of $L^{\infty}_{weak-star}((0,T); L^2_{weak}(\Omega)^3)$ and $\boldsymbol{u}$ is smooth, then, in fact, $(\boldsymbol{u}, \boldsymbol{u})$ will be in the smaller set, which is the closure of the graph for the strong topology of $L^{\infty}((0,T); L^2(\Omega)^3)$.

At this stage the possible relations between Young-measure solutions, dissipative solutions, or our very weak solutions remain an open question.

2. Main results.

Let $0 < T < \infty$. We assume that Ω is a smooth bounded domain in $\mathbf{R}^3$, and $\Omega_t = \{t\} \times \Omega$. We will denote by $\boldsymbol{\nu}$ the outward unit normal vector to Ω.

We introduce the spaces $\mathcal{V}$ and V defined by

$$\mathcal{V} = \left\{ \boldsymbol{v} \in L^2(\Omega)^3 : \frac{\partial v_i}{\partial x_i} = 0 \text{ in } \Omega,\ \boldsymbol{v} \cdot \boldsymbol{\nu} = 0 \text{ on } \partial\Omega \right\} \tag{2.1}$$

and

$$V = L^{\infty}((0,T); \mathcal{V}). \tag{2.2}$$

Let $\boldsymbol{v}^0 \in \mathcal{V}$. We recall the definition of a weak solution.

Definition 1. *A* $\boldsymbol{v} \in V$ *is a weak solution to* (1.1)–(1.4) *with* $\rho = 1$ *if*

$$-\int\int_Q v_i \frac{\partial \phi_i}{\partial t}\, dt d\boldsymbol{x} - \int\int_Q v_j v_i \frac{\partial \phi_i}{\partial x_j}\, dt d\boldsymbol{x} - \int_{\Omega} v_i^0 \phi_i \, d\boldsymbol{x} = 0 \tag{2.3}$$

for all $\boldsymbol{\phi} \in C^1([0,T]; W^{3,2}(\Omega)^3) \cap V$ *with* $\boldsymbol{\phi}(T, \cdot\,) = 0$, *where* $Q = (0,T) \times \Omega$.

Let $\boldsymbol{u} \in V$ be given. We are looking for a function $\boldsymbol{v} \in V$ such that for any $\boldsymbol{\phi}$ as described above

$$-\int\int_Q v_i \frac{\partial \phi_i}{\partial t}\, dt d\boldsymbol{x} - \int\int_Q u_j v_i \frac{\partial \phi_i}{\partial x_j}\, dt d\boldsymbol{x} - \int_{\Omega} v_i^0 \phi_i \, d\boldsymbol{x} = 0. \tag{2.4}$$

Proposition 1. *For any given* $\boldsymbol{u} \in V$ *and any given* $\boldsymbol{v}_0 \in \mathcal{V}$ *there exists a weak solution to Oseen equation* (2.4).

Proof. We will use the Galerkin method to show existence. Since $\mathcal{V}$ is separable, there exists a basis $\{\boldsymbol{w}^k\}_{k \geq 1}$ of $\mathcal{V}$. It is well known that $\{\boldsymbol{w}^k\}_{k \geq 1}$ can be chosen to be an orthonormal basis in $\mathcal{V}$. Also, since $\mathcal{D}(\Omega)$ is dense in $\mathcal{V}$, we can assume that this basis is made of smooth functions (C^{∞}, for example; cf. [3]).

We look for a function $\boldsymbol{v}^n(t,\boldsymbol{x}) = \sum_{1\le k\le n} a_k(t)\boldsymbol{w}^k(\boldsymbol{x})$, such that $a_k(\cdot) \in C^1([0,T])$, and

$$\int_{\Omega_t} \frac{\partial v_i^n}{\partial t} w_i^k \, d\boldsymbol{x} + \int_{\Omega_t} u_j \frac{\partial v_i^n}{\partial x_j} w_i^k \, d\boldsymbol{x} = 0 \tag{2.5}$$

and

$$\int_\Omega v_i^n(0) w_i^k \, d\boldsymbol{x} = \int_\Omega v_i^0 w_i^k \, d\boldsymbol{x} \tag{2.6}$$

for all $1 \le k \le n$.

Putting $\boldsymbol{v}^n$ in (2.5), we get

$$\frac{1}{2}\int_\Omega \frac{\partial}{\partial t}|\boldsymbol{v}^n|^2 d\boldsymbol{x} = 0; \tag{2.7}$$

hence

$$\frac{1}{2}\int_\Omega |\boldsymbol{v}^n|^2 \, d\boldsymbol{x} = \frac{1}{2}\int_\Omega |\boldsymbol{v}^n(0)|^2 \, d\boldsymbol{x}. \tag{2.8}$$

Since the functions $\sum_{1\le k\le m} b_k(t)\boldsymbol{w}^k(x)$, with $b_k(t) \in C^1([0,T])$ and $b_k(T) = 0$, are dense in $C^1([0,T]; W^{3,2}(\Omega)^3) \cap V$ (with $\boldsymbol{\phi}(T,\cdot) = 0$), the $L^\infty_{weak-star}((0,T); L^2_{weak}(\Omega)^3)$ limit $\boldsymbol{v}$ of (a subsequence of) $\boldsymbol{v}^n$ satisfies (2.4) and $\operatorname{div}\boldsymbol{v} = 0$. We also note that $\boldsymbol{v}^n \cdot \boldsymbol{\nu} = 0$ on the boundary Ω_t for all $t \ge 0$. Since $\boldsymbol{v}^n$ converges in the topology of $L^\infty_{weak-star}((0,T); L^2_{weak}(\Omega)^3)$, it follows by the trace theorem that $\boldsymbol{v}\cdot\boldsymbol{\nu} = \lim_{n\to\infty} \boldsymbol{v}^n \cdot \boldsymbol{\nu} = 0$, where the convergence is taking place within the topology of $L^\infty_{weak-star}((0,T); W^{-1/2,2}_{weak}(\partial\Omega)^3)$ (e.g., see Temam [7, p. 9]). Therefore, $\boldsymbol{v} \in V$, and so $\boldsymbol{v}$ is a weak solution to (2.4). □

REMARK 1. *It is easy to prove that, if* $\boldsymbol{u} \in L^\infty((0,T); W^{3,2}(\Omega)^3)$*, then so is* $\boldsymbol{v}$*, and the solution is unique, assuming that* $\boldsymbol{v}^0 \in W^{3,2}(\Omega)^3$.

DEFINITION 2. *Let* $\boldsymbol{u}^n \in V$ *be given, and let* $\boldsymbol{v}^n \in V$ *be the corresponding solution to the Oseen problem* (2.4). *If* $\boldsymbol{u}^n$ *converges to a function* $\boldsymbol{u}$ *and* $\boldsymbol{v}^n$ *converges to a function* $\boldsymbol{v}$ *in the topology of* $L^\infty_{weak-star}((0,T); L^2_{weak}(\Omega)^3)$ *and* $\boldsymbol{u} = \boldsymbol{v}$*, then* $\boldsymbol{u}$ *is called a very weak solution to the problem* (1.1)–(1.4).

REMARK 2. *Obviously, a weak solution is a very weak solution.*

2.1. Existence of very weak solutions to the Dirichlet problem for Euler's equations on bounded domains.

Let Ω be a bounded domain in $\mathbf{R}^3$ with smooth boundary. Let $\omega \in \mathcal{D}(\mathbf{R}^3)$, $\omega(\boldsymbol{x}) \ge 0$, $\int \omega(\boldsymbol{x})\, d\boldsymbol{x} = 1$, $\omega(\boldsymbol{x}) = \omega(-\boldsymbol{x})$. This is a classical mollifier whose existence can be found in most textbooks. We define $\omega_h(\boldsymbol{x}) := h^{-3}\omega(\boldsymbol{x}/h)$, $h > 0$, and

$$\boldsymbol{u}^h(t,\boldsymbol{x}) := (\omega_h * \boldsymbol{u}(t,\cdot))(\boldsymbol{x}) = \frac{1}{h^3}\int_\Omega \omega\left(\frac{\boldsymbol{x}-\boldsymbol{y}}{h}\right)\boldsymbol{u}(t,\boldsymbol{y})\, d\boldsymbol{y}. \tag{2.9}$$

Clearly, if $\boldsymbol{u}$ is divergence-free, then so is $\boldsymbol{u}^h$. Also, $\int_\Omega |\boldsymbol{u}^h|^2 d\boldsymbol{x} \le c\int_\Omega |\boldsymbol{u}|^2 d\boldsymbol{x}$, where c is a constant independent of h.

We choose the same basis, $\{\boldsymbol{w}^k\}_{k\ge 1}$, of $\mathcal{V}$ as before. They are orthogonal functions in $L^2(\Omega)^3$, and they are smooth (cf. [3]). We denote by $\mathcal{P}$ the L^2-orthogonal projection onto $\mathcal{V}$, $E_n := span\{\boldsymbol{w}^1, \ldots, \boldsymbol{w}^n\}$, and we denote by $\mathcal{P}_n$ the L^2-orthogonal projection onto E_n.

Theorem 1. *Let Ω be a bounded domain in $\mathbf{R}^3$ with smooth boundary. For any $\boldsymbol{v}^0 \in \mathcal{V}$ there exists a very weak solution to the problem* (1.1)–(1.4).

Proof. For a fixed $h > 0$, let us first look for a solution to the problem $\boldsymbol{v} \in V$, and

$$-\int_Q v_i \frac{\partial \phi_i}{\partial t}\, dt d\boldsymbol{x} - \int_Q \mathcal{P}(\boldsymbol{v}^h)_j v_i \frac{\partial \phi_i}{\partial x_j}\, dt d\boldsymbol{x} - \int_\Omega v_i^0 \phi_i \, d\boldsymbol{x} = 0 \tag{2.10}$$

for all $\boldsymbol{\phi} \in C^1([0,T]; W^{3,2}(\Omega)^3) \cap V$ with $\boldsymbol{\phi}(T,\cdot) = 0$. Here $\mathcal{P}(\boldsymbol{v}^h)(t,\cdot)$ means $\mathcal{P}(\boldsymbol{v}^h(t,\cdot))$.

As in the above, we will use the Galerkin method to show the existence of solutions to (2.10). Let $\boldsymbol{v}^n$ be the Galerkin approximation to (2.10). That is, $\boldsymbol{v}^n(t,\boldsymbol{x}) = \sum_{1\le k\le n} a_k(t)\boldsymbol{w}^k(\boldsymbol{x}) \in E_n$ with $a_k(\cdot) \in C^1([0,T])$, and $\boldsymbol{v}^n$ satisfies

$$\int_{\Omega_t} \frac{\partial v_i^n}{\partial t} w_i^k \, d\boldsymbol{x} + \int_{\Omega_t} \mathcal{P}(\boldsymbol{v}^{h,n})_j \frac{\partial v_i^n}{\partial x_j} w_i^k \, d\boldsymbol{x} = 0 \tag{2.11}$$

and

$$\int_\Omega v_i^n(0) w_i^k \, d\boldsymbol{x} = \int_\Omega v_i^0 w_i^k \, d\boldsymbol{x} \tag{2.12}$$

for all $1 \le k \le n$. Since $\mathcal{P}(\boldsymbol{v}^{h,n})$ is linear in $a_1, a_2, \dots, a_n$, the second term in the left-hand side of (2.11) is quadratic in $a_1, a_2, \dots, a_n$. Therefore, the existence and uniqueness of $\boldsymbol{v}^n$ follows immediately from the Picard theorem. Clearly, we have

$$\int_\Omega |\boldsymbol{v}^n(t,\boldsymbol{x})|^2 \, d\boldsymbol{x} \le c < \infty \qquad \forall t \in [0,T]. \tag{2.13}$$

We remark that

$$\boldsymbol{v}^{h,n}(t,\boldsymbol{x}) := \frac{1}{h^3} \int_{\mathbf{R}^3} \omega\left(\frac{\boldsymbol{x}-\boldsymbol{y}}{h}\right) \boldsymbol{v}^n(t,\boldsymbol{y}) d\boldsymbol{y}, \qquad \boldsymbol{x} \in \mathbf{R}^3, \tag{2.14}$$

where $\boldsymbol{v}^n(t,\cdot)$ is extended by 0 outside of Ω. From (2.13) and properties of the mollifiers, it follows that $\boldsymbol{v}^{h,n}(t,\cdot) \in C^\infty(\mathbf{R}^3)^3$, and for all $r \ge 0$

$$||\boldsymbol{v}^{h,n}(t,\cdot)||_{C^r(\mathbf{R}^3)} \le c < \infty, \quad t \in [0,T], \tag{2.15}$$

where $c = c(r)$ is a constant.

In order to estimate $\frac{\partial \boldsymbol{v}^n}{\partial t}$, we need the following technical lemma.

Lemma 1.[1] *Let $s \ge 1$ and $p \ge 2$. Assume that $\partial\Omega$ is of class C^∞. Every function $\boldsymbol{\phi} \in W^{s,p}(\Omega)^3$ has the $L^2(\Omega)^3$-orthogonal decomposition*

$$\boldsymbol{\phi} = \boldsymbol{\psi} + \mathbf{grad}(q), \tag{2.16}$$

where $\boldsymbol{\psi} \in W^{s,p}(\Omega)^3 \cap \mathcal{V}$, $q \in W^{s+1,p}(\Omega)/R$. Moreover,

$$||\boldsymbol{\psi}||_{W^{s,p}(\Omega)^3} \le C\, ||\boldsymbol{\phi}||_{W^{s,p}(\Omega)^3}, \tag{2.17}$$

where $C = C(s,p,\Omega)$ is a constant independent of $\boldsymbol{\phi}$.

[1] This should be a well-known classical result about the Helmholtz decomposition. But since we were unable to find a easy source to reference, we decided to provide a quick proof for the reader's convenience.

Proof of Lemma 1. Let $\boldsymbol{\phi} \in W^{s,p}(\Omega)^3$. As $p \geq 2$, $\boldsymbol{\phi} \in L^2(\Omega)^3$, and the Helmholtz decomposition yields (2.16), where $\boldsymbol{\psi} \in L^2(\Omega)^3$, $\operatorname{div}\boldsymbol{\psi} = 0$ in Ω, $\boldsymbol{\psi}\cdot\boldsymbol{\nu} = 0$ on $\partial\Omega$ (i.e., $\boldsymbol{\psi} \in \mathcal{V}$), $q \in W^{1,2}(\Omega)/R$, and $\boldsymbol{\psi}$, $\mathbf{grad}(q)$ are mutually orthogonal in $L^2(\Omega)^3$ (cf. [7, Theorem 1.4., p. 15] or [2, Corollary 3.4, p. 50]). As $\boldsymbol{\phi} \in W^{s,p}(\Omega)^3$, with $s \geq 1$, we have $\boldsymbol{\phi}\cdot\boldsymbol{\nu} \in W^{s-1/p,p}(\partial\Omega)$. It then follows that $\Delta q = \operatorname{div}(\boldsymbol{\phi}-\boldsymbol{\psi}) = \operatorname{div}\boldsymbol{\phi} \in W^{s-1,p}(\Omega)$ and $\frac{\partial q}{\partial \boldsymbol{\nu}} = \mathbf{grad}(q)\cdot\boldsymbol{\nu} = (\boldsymbol{\phi}-\boldsymbol{\psi})\cdot\boldsymbol{\nu} = \boldsymbol{\phi}\cdot\boldsymbol{\nu} \in W^{s-1/p,p}(\partial\Omega)$. The regularity of the solution to the Neumann problem for the Laplace equation implies (cf. [2, Theorem 1.10, p. 15]) that $q \in W^{s+1,p}(\Omega)/R$ and there is a constant $C = C(s,p,\Omega)$ (independent of $\boldsymbol{\phi}$) such that $||q||_{W^{s+1,p}(\Omega)^3/R} \leq C\left(||\operatorname{div}\boldsymbol{\phi}||_{W^{s-1,p}(\Omega)} + ||\boldsymbol{\phi}\cdot\boldsymbol{\nu}||_{W^{s-1/p,p}(\partial\Omega)}\right) \leq C\,||\boldsymbol{\phi}||_{W^{s,p}(\Omega)^3}$. Thus $\boldsymbol{\psi} = \boldsymbol{\phi} - \mathbf{grad}(q) \in W^{s,p}(\Omega)^3$ and $||\boldsymbol{\psi}||_{W^{s,p}(\Omega)^3} \leq ||\boldsymbol{\phi}||_{W^{s,p}(\Omega)^3} + ||q||_{W^{s+1,p}(\Omega)^3/R}$, which yields (2.17). □

COROLLARY 1. *If $s, p, \Omega, \boldsymbol{\phi}$, and $\boldsymbol{\psi}$ are as in Lemma* 1, *then $\boldsymbol{\psi} = \mathcal{P}(\boldsymbol{\phi})$. Therefore, the operator $\mathcal{P}$ maps $W^{s,p}(\Omega)^3$ onto $W^{s,p}(\Omega)^3 \cap \mathcal{V}$, and $\mathcal{P}$ is a bounded operator.*

Proof of Corollary 1. In (2.16) $\boldsymbol{\psi} \in \mathcal{V}$ and $\mathbf{grad}(q)$ is in the $L^2(\Omega)^3$-orthogonal of $\mathcal{V}$, so $\mathcal{P}(\boldsymbol{\phi}) = \boldsymbol{\psi}$. Equation (2.17) implies that $\mathcal{P}$ is a bounded operator. □

Proof of Theorem 1 (cont.). Let $\boldsymbol{\phi} \in W_0^{1,2}(\Omega)^3$ and $\boldsymbol{\psi}$ be as in (2.16). That is, $\boldsymbol{\psi} = \mathcal{P}(\boldsymbol{\phi})$. Since $\boldsymbol{v}^n(t,\cdot) \in E_n$, and so $\frac{\partial \boldsymbol{v}^n}{\partial t}(t,\cdot) \in E_n$, using (2.11), we have

$$\begin{aligned}\int_{\Omega_t} \frac{\partial v_i^n}{\partial t}\phi_i\,d\boldsymbol{x} = \int_{\Omega_t} \frac{\partial v_i^n}{\partial t}\psi_i\,d\boldsymbol{x} \;&=\; \int_{\Omega_t} \frac{\partial v_i^n}{\partial t}\mathcal{P}_n(\boldsymbol{\psi})_i\,d\boldsymbol{x}\\ &= \int_{\Omega_t} \mathcal{P}(v^{h,n})_j v_i^n \frac{\partial}{\partial x_j}\left(\mathcal{P}_n\left(\boldsymbol{\psi}\right)_i\right)\,d\boldsymbol{x}.\end{aligned} \tag{2.18}$$

As $\boldsymbol{v}^{h,n} \in C^\infty(\mathbf{R}^3)$, then $\boldsymbol{v}^{h,n} \in W^{1,4}(\Omega)$. The embedding theorem, Corollary 1, (2.15), and the properties of mollifiers imply that

$$\begin{aligned}\left|\left|\mathcal{P}(\boldsymbol{v}^{h,n})\right|\right|_{L^\infty(\Omega)^3} &\leq c\,\left|\left|\mathcal{P}(\boldsymbol{v}^{h,n})\right|\right|_{W^{1,4}(\Omega)^3}\\ &\leq c\,\left|\left|\boldsymbol{v}^{h,n}\right|\right|_{W^{1,4}(\Omega)^3} \leq c\,||\boldsymbol{v}^n||_{L^2(\Omega)^3} \leq c.\end{aligned} \tag{2.19}$$

Since $E_n \subset \mathcal{V}$, it follows that $\mathcal{P}_n(\boldsymbol{\psi}) = \mathcal{P}_n(\boldsymbol{\phi})$. However, $\boldsymbol{\phi} \in W_0^{1,2}(\Omega)^3$, and (for an appropriate choice of the basis $\{\boldsymbol{w}^k\}_{k\geq 1}$ of $\mathcal{V}$) a direct calculation yields that $\mathcal{P}_n(\boldsymbol{\phi})$ is also the $W^{1,2}(\Omega)^3$-orthogonal projection of $\boldsymbol{\phi}$ on E_n. Therefore,

$$||\mathcal{P}_n(\boldsymbol{\psi})||_{W^{1,2}(\Omega)^3} = ||\mathcal{P}_n(\boldsymbol{\phi})||_{W^{1,2}(\Omega)^3} \leq c\,||\boldsymbol{\phi}||_{W_0^{1,2}(\Omega)^3}\,. \tag{2.20}$$

Using (2.19), (2.13), and (2.20) in (2.18), we obtain

$$\left|\int_{\Omega_t} \frac{\partial v_i^n}{\partial t}\phi_i\,d\boldsymbol{x}\right| \leq\; c\,||\boldsymbol{\phi}||_{W_0^{1,2}(\Omega)^3} \tag{2.21}$$

for any $\boldsymbol{\phi} \in W_0^{1,2}(\Omega)^3$. Thus

$$\left|\left|\frac{\partial \boldsymbol{v}^n}{\partial t}\right|\right|_{L^\infty((0,T);W^{-1,2}(\Omega)^3)} \leq\; c. \tag{2.22}$$

From (2.14) and (2.22), it follows from classical properties of mollifiers that

$$\left|\left|\frac{\partial \boldsymbol{v}^{h,n}}{\partial t}\right|\right|_{L^\infty((0,T);W^{1,2}(\Omega)^3)} \leq\; c. \tag{2.23}$$

Note that in (2.19)–(2.23) the constant c depends on h.

Let $h > 0$ be fixed. From (2.13) and classical compactness arguments it follows that there exist a subsequence n_k and a function $\boldsymbol{v} \in V$ such that $\boldsymbol{v} = \lim_{n_k \to \infty} \boldsymbol{v}^{n_k}$ in the topology of $L^\infty_{weak-star}((0,T); L^2_{weak}(\Omega)^3)$. It then follows from (2.23) and (2.15) that $\boldsymbol{v}^{h,n_k}$ converges strongly in $L^2((0,T); L^2(\Omega)^3)$ to $\boldsymbol{v}^h$ as $k \to \infty$, and so $\mathcal{P}(\boldsymbol{v}^{h,n_k}) \to \mathcal{P}(\boldsymbol{v}^h)$ strongly in $L^2((0,T); L^2(\Omega)^3)$. Therefore, $\mathcal{P}(\boldsymbol{v}^{h,n_k})_j v_i^{n_k}$ converges in the topology of $L^\infty_{weak-star}((0,T); L^2_{weak}(\Omega))$ to $\mathcal{P}(\boldsymbol{v}^h)_j v_i$. Hence $\boldsymbol{v}$ is a solution to the problem (2.10).

We will now let h go to zero and finish the proof of the theorem.

Let us consider a subsequence $h_m \to 0$, and let $\boldsymbol{v}^m$ be the corresponding solution to (2.10) with $h = h_m$. It can easily be seen that $\boldsymbol{v}^m$ and $\boldsymbol{v}^{h_m,m}$ are uniformly bounded in $L^\infty((0,T); L^2(\Omega))$. Therefore, without loss of generality, we can assume that $\boldsymbol{v}^m \to \boldsymbol{v}$ and $\boldsymbol{v}^{h_m,m} \to \boldsymbol{w}$, and so $\mathcal{P}(\boldsymbol{v}^{h_m,m}) \to \mathcal{P}(\boldsymbol{w})$, in the topology of $L^\infty_{weak-star}((0,T); L^2_{weak}(\Omega)^3)$. Clearly, we will have that $\boldsymbol{v} \in V$. To finish the proof of the theorem we need simply to show that $\mathcal{P}(\boldsymbol{w}) = \boldsymbol{v}$.

First we will show that $\boldsymbol{w} = \boldsymbol{v}$. Letting $\boldsymbol{\chi} \in C(\bar{I}; L^2(\Omega)^3)$, we have

$$\begin{aligned}
&\int_0^T \int_\Omega v_i^{h_m,m}(t,\boldsymbol{x})\chi_i(t,\boldsymbol{x})\, d\boldsymbol{x}\, dt \\
&\qquad = \int_0^T dt \int_\Omega \chi_i(t,\boldsymbol{x}) d\boldsymbol{x} \int_{\mathbf{R}^3} \omega\left(\frac{\boldsymbol{x}-\boldsymbol{y}}{h_m}\right) v_i^m(t,\boldsymbol{y})\, d\boldsymbol{y} \\
&\qquad = \int_0^T dt \int_{\mathbf{R}^3} v_i^m(t,\boldsymbol{y}) \frac{d\boldsymbol{y}}{h_m^3} \int_{\mathbf{R}^3} \omega\left(\frac{\boldsymbol{x}-\boldsymbol{y}}{h_m}\right) \chi_i(t,\boldsymbol{x})\, d\boldsymbol{x}
\end{aligned} \tag{2.24}$$

(with $\chi_i(t,\cdot) = 0$ on $\mathbf{R}^3 \setminus \Omega$). Since

$$\frac{1}{h^3} \int_{\mathbf{R}^3} \omega\left(\frac{\boldsymbol{y}-\boldsymbol{x}}{h}\right) \chi_i(t,\boldsymbol{x})\, d\boldsymbol{x} \to \chi_i(t,\boldsymbol{y}) \qquad \text{in } C(\bar{I} \times \Omega)$$

as $h \to 0$ and $\omega(\boldsymbol{x}) = \omega(-\boldsymbol{x})$, we get from (2.24) that

$$\int_0^T \int_\Omega v_i^{h_m,m}(t,\boldsymbol{x})\chi_i(t,\boldsymbol{x})\, d\boldsymbol{x}\, dt \to \int_0^T \int_\Omega v_i(t,\boldsymbol{y})\chi_i(t,\boldsymbol{y})\, d\boldsymbol{y}\, dt. \tag{2.25}$$

It follows that $\boldsymbol{w} = \boldsymbol{v}$.

Since $\boldsymbol{v} \in V$, it follows that $\boldsymbol{w} \in V$, and therefore $\mathcal{P}(\boldsymbol{w}) = \boldsymbol{w}$. Hence $\mathcal{P}(\boldsymbol{w}) = \boldsymbol{v}$, and so $\boldsymbol{v}$ is a very weak solution. □

2.2. Existence of space-periodic very weak solutions to Euler's equation. We shall now consider the space-periodic solutions to (1.1)–(1.4). The proof here is similar to that of the previous subsection. But since in the periodic case we can work with an explicit basis and an explicit decomposition (2.16), we will present some of the details.

We set $\Omega = (-\pi,\pi)^3$ and $Q = (0,T) \times (-\pi,\pi)^3$. There is an obvious identification between functions on Ω and the functions 2π-periodic in each variable on $\mathbf{R}^3$. Throughout this subsection we will use this identification. We consider the spaces $\mathcal{V}_{per}$ and V_{per} defined by

$$\mathcal{V}_{per} = \mathcal{V}_{per}(\mathbf{R}^3) = \left\{ \boldsymbol{v} \in L^2((-\pi,\pi)^3)^3 : \frac{\partial v_i}{\partial x_i} = 0 \text{ in } (-\pi,\pi)^3,\ \int_{(-\pi,\pi)^3} \boldsymbol{v}\, d\boldsymbol{x} = 0 \right\}$$

and

$$V_{per} = L^\infty((0,T); \mathcal{V}_{per}). \tag{2.26}$$

We set $\boldsymbol{w}^{\boldsymbol{k}} = \boldsymbol{c}_{\boldsymbol{k}} e^{I(\boldsymbol{x}\cdot\boldsymbol{k})} + \bar{\boldsymbol{c}}_{\boldsymbol{k}} e^{-I(\boldsymbol{x}\cdot\boldsymbol{k})}$ with $c^j_{\boldsymbol{k}} k_j = 0$, $\boldsymbol{k} \neq (0,0,0)$, where $I^2 = -1$, $\bar{\boldsymbol{c}}_{\boldsymbol{k}} = (\bar{c}^1_{\boldsymbol{k}}, \bar{c}^2_{\boldsymbol{k}}, \bar{c}^3_{\boldsymbol{k}})$, and $\bar{c}^j_{\boldsymbol{k}}$ is the complex conjugate of $c^j_{\boldsymbol{k}}$.

Thus $\{\boldsymbol{w}^{\boldsymbol{k}}\}_{|\boldsymbol{k}|>0}$ is an orthogonal basis in $\mathcal{V}_{per}$, where $|\boldsymbol{k}| = \max_i |k_i|$. We denote $E_n := span\{\boldsymbol{w}^{\boldsymbol{k}} : 0 < |\boldsymbol{k}| \le n\}$.

A function $\boldsymbol{v} \in V_{per}$ is a *space-periodic weak solution* to (1.1)–(1.4), with $\rho = 1$, if $\boldsymbol{v}$ satisfies (2.3) for all $\boldsymbol{\phi}$ as above with $\boldsymbol{\phi}$ 2π-periodic in each spatial variable.

PROPOSITION 2. *For any $\boldsymbol{u} \in V_{per}$ and any given $\boldsymbol{v}^0 \in \mathcal{V}_{per}$, there exists a space-periodic weak solution to Oseen equation* (2.4).

Proof. The proof is similar to that of Proposition 1. □

A function $\boldsymbol{u} \in V_{per}$ is called a *space-periodic very weak solution* to the problem (1.1), (1.2), and (1.4) if Definition 2 holds for $\boldsymbol{u}$ with $\boldsymbol{u}^n, \boldsymbol{v}^n \in V_{per}$.

THEOREM 2. *For any $\boldsymbol{v}^0 \in \mathcal{V}_{per}$ there exists a space-periodic very weak solution to the problem* (1.1)*,* (1.2)*, and* (1.4).

Proof. As in the proof of Theorem 1, for a fixed $h > 0$, we consider the equation for $\boldsymbol{v} \in V_{per}$

$$-\int_Q v_i \frac{\partial \phi_i}{\partial t}\, dt\, d\boldsymbol{x} - \int_Q v^h_j v_i \frac{\partial \phi_i}{\partial x_j}\, dt\, d\boldsymbol{x} - \int_\Omega v^0_i \phi_i\, d\boldsymbol{x} = 0 \tag{2.27}$$

for all $\boldsymbol{\phi} \in C^1([0,T]; W^{3,2}(\Omega)^3) \cap V_{per}$, $\boldsymbol{\phi}$ 2π-periodic in each spatial variable, with $\boldsymbol{\phi}(T, \cdot\,) = 0$. Note that $\boldsymbol{v}^h(t, \cdot\,) \in C^\infty(\mathbf{R}^3) \cap \mathcal{V}_{per}$, 2π-periodic function in each spatial variable, with $\int_{\Omega_t} \boldsymbol{v}^h\, d\boldsymbol{x} = 0$, for all t. That is, $\boldsymbol{v}^h \in V_{per}$.

As above, let $\boldsymbol{v}^n(t,x) = \sum_{0<|\boldsymbol{k}|\le n} a_{\boldsymbol{k}}(t) \boldsymbol{w}^{\boldsymbol{k}}(x)$ be the Galerkin approximation in E_n to (2.27), and let $\boldsymbol{v}^{h,n}$ be defined by (2.14), where $\boldsymbol{v}^n$ is considered as a 2π-periodic function on $\mathbf{R}^3$. Clearly, $\boldsymbol{v}^{h,n} \in C^\infty(\mathbf{R}^3)^3$, and $\boldsymbol{v}^{h,n}$ is 2π-periodic in each spatial variable. The function $\boldsymbol{v}^n$ satisfies

$$\int_{\Omega_t} \frac{\partial v^n_i}{\partial t} w^{\boldsymbol{k}}_i\, d\boldsymbol{x} + \int_{\Omega_t} v^{h,n}_j \frac{\partial v^n_i}{\partial x_j} w^{\boldsymbol{k}}_i\, d\boldsymbol{x} = 0 \tag{2.28}$$

and

$$\int_\Omega v^n_i(0) w^{\boldsymbol{k}}_i\, d\boldsymbol{x} = \int_\Omega v^0_i w^{\boldsymbol{k}}_i\, d\boldsymbol{x} \tag{2.29}$$

for all $1 \le k \le n$. The existence and uniqueness of such a function follows immediately from the Picard theorem. We have

$$||\boldsymbol{v}^n||_{L^\infty((0,T), L^2(\Omega)^3)} \le c. \tag{2.30}$$

Let $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3) \in W^{1,2}(\Omega)^3$ be a 2π-periodic function in each variable. Thus $\phi_i(\boldsymbol{x}) = \sum_{\boldsymbol{\ell}} \phi_{i,\boldsymbol{\ell}} e^{I(\boldsymbol{\ell},\boldsymbol{x})}$ with $\sum_{\boldsymbol{\ell}} |\phi_{i,\boldsymbol{\ell}}|^2 |\boldsymbol{\ell}|^2_2 < \infty$ and $|\boldsymbol{\ell}|^2_2 = \ell^2_1 + \ell^3_1 + \ell^2_3$. We define

$$q_{\boldsymbol{\ell}} = -I \frac{\ell_i\, \phi_{i,\boldsymbol{\ell}}}{|\boldsymbol{\ell}|^2_2}, \quad \psi_{i,\boldsymbol{\ell}} = \phi_{i,\boldsymbol{\ell}} - I \ell_i\, q_{\boldsymbol{\ell}}$$

and set $\psi_i(\boldsymbol{x}) = \sum_{\boldsymbol{\ell}} \psi_{i,\boldsymbol{\ell}} e^{I(\boldsymbol{\ell},\boldsymbol{x})}$, $\boldsymbol{\psi} = (\psi_1, \psi_2, \psi_3)$, and $q(\boldsymbol{x}) = \sum_{\boldsymbol{\ell}} q_{\boldsymbol{\ell}} e^{I(\boldsymbol{\ell},\boldsymbol{x})}$. It is easy to verify that $\boldsymbol{\psi} \in W^{1,2}(\Omega)^3$, $\operatorname{div} \boldsymbol{\psi} = 0$ in Ω, $||\mathbf{grad}(q)||_{W^{1,2}(\Omega)^3} \le c||\boldsymbol{\phi}||_{W^{1,2}(\Omega)^3}$,

$\boldsymbol{\psi}, \mathbf{grad}(q)$ are orthogonal in both $L^2(\Omega)^3$ and $\boldsymbol{\phi} = \boldsymbol{\psi} + \mathbf{grad}(q)$. This is exactly the analogue of (2.16). Using this decomposition and (2.28), we immediately get (2.21) and then

$$\left|\left|\frac{\partial \boldsymbol{v}^n}{\partial t}\right|\right|_{L^\infty((0,T);W^{-1,2}(\Omega)^3)} \leq c. \tag{2.31}$$

Let $h > 0$ be fixed. It follows from (2.30) that there exist a subsequence $\boldsymbol{v}^{n_k}$ and a function $\boldsymbol{v} \in V_{per}$ such that $\boldsymbol{v}^{n_k} \to \boldsymbol{v}$ in $L^\infty_{weak-star}((0,T); L^2_{weak}(\Omega)^3)$. Since $L^2(\Omega)^3 \subset\subset W^{-1,2}(\Omega)^3 \subset W^{-1,2}(\Omega)^3$, using (2.30) and (2.31), the convergence $\boldsymbol{v}^{n_k} \to \boldsymbol{v}$ holds in $L^2((0,T); W^{-1,2}(\Omega)^3)$ (cf. [3, Theorem 5.1, p. 58] or [7, Theorem 2.1, p. 271]). Using properties of mollifiers, it then follows that $\boldsymbol{v}^{h,n_k} \to \boldsymbol{v}^h$ strongly in $L^2((0,T); W^{2,2}(\Omega)^3)$. For a function $\boldsymbol{\chi} \in L^2((0,T); W^{2,2}(\Omega)^3)$, $v_j^{h,n_k} \frac{\partial \chi_i}{\partial x_j} \to v_j^h \frac{\partial \chi_i}{\partial x_j}$ in $L^2((0,T); W^{1,2}(\Omega))$, and so $\int_0^T \int_\Omega v_j^{h,n_k} v_i^{n_k} \frac{\partial \chi_i}{\partial x_j}\, d\boldsymbol{x}dt \to \int_0^T \int_\Omega v_j^h v_i \frac{\partial \chi_i}{\partial x_j}\, d\boldsymbol{x}dt$. Therefore, $\boldsymbol{v}$ is a solution to the problem (2.27).

Proceeding as in the last part of the proof of Theorem 1, by letting h go to zero, we obtain (2.25). Therefore, $\boldsymbol{v}$ is a space-periodic very weak solution, and the proof is complete. □

3. Regularity. Here we state some regularity results and shed some light on the relation between very weak solutions and traditional classical solutions. The first result we prove is that if a smooth solution exists, which implicitly assumes that the initial condition is smooth, then the very weak solution constructed by our method will coincide with the classical solution. This does not preclude the existence of other, nonsmooth, very weak solutions.

The second theorem we prove in this section is less obvious and is more significant. We prove that *any* smooth very weak solution is actually a classical solution. This is achieved by showing that smooth elements of the form $(\boldsymbol{u}, \boldsymbol{u})$, which are in the closure, for the weak topology, of the graph of the map γ described in the introduction are in fact in the *smaller* set, which is the closure for the strong topology of the same set.

For simplicity of notation we prove these results in the space-periodic case. The generalization to the other case is immediate.

We begin by clarifying our terminology.

DEFINITION 3. *A function $\boldsymbol{u} \in V_{per}$ is said to be a smooth classical solution to the problem* (1.1), (1.2), *and* (1.4) *in* $[0,T) \times \Omega$ *if it is in the space* $C^0([0,T); W^{3,2}(\Omega)^3)$ *and it satisfies* (1.1), (1.2), *and* (1.4).

LEMMA 2. *Assume that $\boldsymbol{v}^0 \in \mathcal{V} \cap W^{3,2}(\Omega)^3$ and $h > 0$ is a fixed number. Then there exist a time $T > 0$ and a constant C, both independent of h, such that the solution $\boldsymbol{v}(t, \boldsymbol{x}, h)$ of problem* (2.27) *satisfies*

$$||\boldsymbol{v}(\cdot, \cdot, h)||_{C^0([0,T), W^{3,2}(\Omega)^3)} \leq C. \tag{3.1}$$

Proof. This proof follows the same outline as the existence proof used in [6]. Using $\boldsymbol{\phi} = D^{\boldsymbol{l}} \boldsymbol{v}^n$ for all $|\boldsymbol{l}| \leq 3$ as a test function in (2.28), we deduce as in [6, section 2.1] that

$$\frac{d}{dt} ||\boldsymbol{v}^n(t, \cdot)||^2_{W^{3,2}(\Omega)^3} \leq c ||\boldsymbol{v}^n(t, \cdot)||^3_{W^{3,2}(\Omega)^3}, \tag{3.2}$$

where c is a constant independent of n and h. It then easily follows that by letting $n \to \infty$, we get that the solution $\boldsymbol{v}(t, \boldsymbol{x}, h)$ of problem (2.27) satisfies

$$\frac{d}{dt} ||\boldsymbol{v}(t, \cdot, h)||^2_{W^{3,2}(\Omega)^3} \leq c ||\boldsymbol{v}(t, \cdot, h)||^3_{W^{3,2}(\Omega)^3} \tag{3.3}$$

for all $t \geq 0$ and all h.

Choose T such that $0 < T < \frac{2}{c \cdot ||\boldsymbol{v}^0||_{W^{3,2}(\Omega)^3}}$. From the previous estimate it easily follows (see [6, Appendix 2.1], for example) that for $0 \leq t \leq T$

$$||\boldsymbol{v}(\cdot, \cdot, h)||_{C^0([0,T);W^{3,2}(\Omega)^3)} \leq \frac{2||\boldsymbol{v}^0||_{W^{3,2}(\Omega)^3}}{2 - T \cdot c \cdot ||\boldsymbol{v}^0||_{W^{3,2}(\Omega)^3}}, \tag{3.4}$$

where c (the same constant as in the estimate (3.3)) is independent of h and of the initial condition. □

PROPOSITION 3. *Assume that $\boldsymbol{v}^0 \in \mathcal{V}_{per} \cap W^{3,2}(\Omega)^3$. Then there exists a time $T > 0$ such that the space-periodic very weak solution $\boldsymbol{v}(t, \boldsymbol{x})$ to problem* (1.1), (1.2), (1.4) *obtained as a limit of solutions $\boldsymbol{v}(t, \boldsymbol{x}, h)$ to the problem* (2.27) *is a smooth classical solution on $[0, T) \times \Omega$, and it satisfies estimates* (3.3)–(3.4).

Proof. Let $T > 0$ be small enough so that the right-hand side of (3.4) is finite. Since the very weak solution $\boldsymbol{v}$ is the weak limit in L^2 of a sequence of functions which satisfy the estimates (3.3)–(3.4) uniformly, it easily follows that this sequence of functions converges strongly to the weak solution $\boldsymbol{v}$. It follows from this and the Oseen equation that such a limit is a weak solution of the Euler system of equations.

From (3.3)–(3.4) it follows that this weak solution satisfies the same estimates, and we can then deduce that such a weak solution is in fact a classical solution on $[0, T) \times \Omega$. □

Next we prove a continuation-type technical lemma.

LEMMA 3. *Let $M > 0$ be a given number, and let $\boldsymbol{v}(t, \boldsymbol{x})$ be the space-periodic very weak solutions to problem* (1.1), (1.2), (1.4) *obtained as a limit of solutions $\boldsymbol{v}(t, \boldsymbol{x}, h)$ to the problem* (2.27). *Assume that $\boldsymbol{v}^0 \in \mathcal{V}_{per} \cap W^{3,2}(\Omega)^3$ and that there exists a finite time $T_M > 0$ such that $||\boldsymbol{v}(t, \cdot)||_{W^{3,2}(\Omega)^3} \leq M$ for all $t \leq T_M$. Then there exists a number $\alpha > 0$ such that $\boldsymbol{v}(t, \boldsymbol{x})$ is a smooth classical solution on $[0, T_M + \alpha) \times \Omega$.*

Proof. Here c is the same constant as in estimate (3.3). We have already proved in the proposition above that this very weak solution is a smooth classical solution in the region $[0, \sigma) \times \Omega$ with $\sigma = \frac{2}{c \cdot ||\boldsymbol{v}^0||_{W^{3,2}(\Omega)^3}}$.

Now we will consider the remaining case $T_M \geq \sigma$. Let $\sigma_{M+1} = \frac{2}{c \cdot (M+1)} > \sigma_{M+2} = \frac{2}{c \cdot (M+2)}$. Assume that $0 < t_0 < T_M$ is such that $\boldsymbol{v}(t, \boldsymbol{x})$ is a smooth classical solution on $[0, t_0] \times \Omega$ and that $\boldsymbol{v}(t_0, \boldsymbol{x}, h)$ converges weakly in $W^{3,2}(\Omega)^3$ to $\boldsymbol{v}(t_0, \boldsymbol{x})$ as $h \to 0$. From the previous proposition it follows that the set of such t_0 is not empty.

We will show that in fact $\boldsymbol{v}(t, \boldsymbol{x})$ is a smooth classical solution on $[0, t_0 + \sigma_{M+2}] \times \Omega$ and that $\boldsymbol{v}(t_0 + \sigma_{M+2}, \boldsymbol{x}, h)$ converges weakly in $W^{3,2}(\Omega)^3$ to $\boldsymbol{v}(t_0 + \sigma_{M+2}, \boldsymbol{x})$.

Since $||\boldsymbol{v}(t_0, \cdot)||_{W^{3,2}(\Omega)^3} \leq M$, it follows from the weak convergence that for h small enough $||\boldsymbol{v}(t_0, \cdot, h)||_{W^{3,2}(\Omega)^3} \leq (M + 1)$. Dividing both sides of (3.3) by $||\boldsymbol{v}(t, \cdot, h)||^3_{W^{3,2}(\Omega)^3}$ and integrating with respect to time from t_0 to $t_0 + s$, we find that

$$||\boldsymbol{v}(t_0 + s, \cdot, h)||_{W^{3,2}(\Omega)^3} \leq \frac{2||\boldsymbol{v}(t_0, \cdot, h)||_{W^{3,2}(\Omega)^3}}{2 - s \cdot c \cdot ||\boldsymbol{v}(t_0, \cdot, h)||_{W^{3,2}(\Omega)^3}}, \tag{3.5}$$

from which it easily follows that

$$\sup_{s \leq \sigma_{M+2}} ||\boldsymbol{v}(t_0 + s, \cdot, h)||_{W^{3,2}(\Omega)^3} \leq \frac{2(M+1)}{2 - \sigma_{M+2} \cdot c \cdot (M+1)}. \tag{3.6}$$

This, in conjunction with (3.3), will allow us to conclude that the sequence $\boldsymbol{v}(t, \boldsymbol{x}, h)$ converges weakly in $W^{1,2}((t_0, t_0 + \sigma_{M+2}); W^{3,2}(\Omega)^3)$ to $\boldsymbol{v}(t, \boldsymbol{x})$. It then holds that

$\boldsymbol{v}(t,\boldsymbol{x})$ is now a smooth classical solution on $[0, t_0 + \sigma_{M+2}] \times \Omega$. We also have that $\boldsymbol{v}(t_0 + \sigma_{M+2}, \boldsymbol{x}, h)$ converges weakly in $W^{3,2}(\Omega)^3$ to $\boldsymbol{v}(t_0 + \sigma_{M+2}, \boldsymbol{x})$.

The procedure described above can be repeated as long as $t_0 + \sigma_{M+2} \leq T_M$. Repeating the previous procedures enough times, we will, in a finite number of steps, reach the time $T_M + \frac{\sigma_{M+2}}{2}$. We can then take $\alpha = \frac{\sigma_{M+2}}{2}$. □

Next we will prove the uniqueness result we alluded to in point 2 of the introduction.

THEOREM 3. *Assume that* $\boldsymbol{v}^0 \in \mathcal{V}_{per} \cap W^{3,2}(\Omega)^3$ *and that* $\boldsymbol{v}(t,\boldsymbol{x})$ *is the space-periodic very weak solution to problem* (1.1)*,* (1.2)*,* (1.4) *obtained as a limit of solutions* $\boldsymbol{v}(t,\boldsymbol{x},h)$ *to the problem* (2.27)*. Assume that there exist a positive time* T_c *and a function* $\boldsymbol{u}$ *which is a smooth classical solution on* $[0,T_c) \times \Omega$ *to the problem* (1.1)*,* (1.2)*, and* (1.4) *in the sense of the Definition* 3*. Assume also that* $\boldsymbol{v}(0,\boldsymbol{x}) = \boldsymbol{u}(0,\boldsymbol{x})$*. Then* $\boldsymbol{v}(t,\boldsymbol{x}) \in C^0([0,T_c); W^{3,2}(\Omega)^3)$*. Moreover,* $\boldsymbol{v}(t,\boldsymbol{x}) = \boldsymbol{u}(t,\boldsymbol{x})$ *for all* $t \in [0,T_c)$*.*

Proof. Let T^* be a positive finite number such that $T^* < T_c$. Assume that $\boldsymbol{v}(t,\boldsymbol{x}) \in C^0([0,T^*]; W^{3,2}(\Omega)^3)$. It then follows from Lemma 3 that $\boldsymbol{v}(t,\boldsymbol{x})$ is a smooth classical solution over the region $[0,T^*] \times \Omega$. In this case, it is an easy exercise (see [6, p. 78], for example) to show that $\boldsymbol{v}(t,\boldsymbol{x}) = \boldsymbol{u}(t,\boldsymbol{x})$ for all $t \in [0, T*)$. If this can be done for all $T^* < T_c$, then the solutions coincide for all $t \in [0,T_c)$.

Alternatively, there exists a $T_b < T_c$ such that $\boldsymbol{v}(t,\boldsymbol{x}) \notin C^0([0,T_b]; W^{3,2}(\Omega)^3)$. In this case it can easily be seen, using Lemma 3, that there exists a positive time $T_{cr} \leq T_b$ such that $\boldsymbol{v}(t,\boldsymbol{x}) \in C^0([0,T_{cr}); W^{3,2}(\Omega)^3)$ and such that

$$\lim_{t\to {T_{cr}}^-} ||\boldsymbol{v}(t,\cdot)||_{W^{3,2}(\Omega)^3} = \infty. \tag{3.7}$$

From the discussion above we will have that $\boldsymbol{v}(t,\boldsymbol{x}) = \boldsymbol{u}(t,\boldsymbol{x})$ for all $t \in [0,T_{cr})$. This with (3.7) would contradict the fact that $\boldsymbol{u}(t,\boldsymbol{x}) \in C^0([0,T_c); W^{3,2}(\Omega)^3)$. □

THEOREM 4. *Let* $\boldsymbol{w} \in V_{per} \cap C^0([0,T_w); W^{3,2}(\Omega)^3)$ *be any space-periodic very weak solution. Then* $\boldsymbol{w}$ *is a classical solution over the time interval* $[0,T_w)$*.*

Proof. Let $\boldsymbol{w}$ be a very weak smooth solution. Then, by definition, there exist a sequence $\boldsymbol{u}^n \in V$ and a sequence $\boldsymbol{v}^n \in V$ such that $\frac{\partial v_i^n}{\partial t} + u_j^n \frac{\partial v_i^n}{\partial x_j} + \frac{\partial p^n}{\partial x_i} = 0$ in $(0,T_w) \times \Omega$, $i = 1,2,3$, and such that $\boldsymbol{v}^n$ and $\boldsymbol{u}^n$ converge to $\boldsymbol{w}$ in $L^\infty_{weak-star}((0,T_w); L^2_{weak}(\Omega)^3)$.

Fix T such that $0 < T < T_w$. We will show next that $\boldsymbol{w}$ is a classical solution in $[0,T) \times \Omega$ in the sense of Definition 3.

Now we introduce the sequence of functions $\boldsymbol{u}^{h,n}$ defined by mollifying the sequence $\boldsymbol{u}^n$ as in (2.14), we let $\boldsymbol{\psi}^{h,n} \in V$ be the unique solution of the Oseen problem obtained by taking $\boldsymbol{u} = \boldsymbol{u}^{h,n}$ in (2.4), and we let the initial condition $\boldsymbol{v}^{h,n}(0,\boldsymbol{x})$ be the mollification of $\boldsymbol{v}^n(0,\boldsymbol{x})$. Taking the difference of the equations satisfied by $\boldsymbol{\psi}^{h,n}$ and $\boldsymbol{v}^n$, multiplying by $\boldsymbol{\psi}^{h,n} - \boldsymbol{v}^n$, and integrating by parts in space, we find

$$\frac{1}{2}\int_\Omega \frac{\partial}{\partial t}|\boldsymbol{\psi}^{h,n} - \boldsymbol{v}^n|^2 d\boldsymbol{x} = \int_\Omega (u^n - u^{h,n})_j \left(\frac{\partial}{\partial x_j}\psi_i^{h,n}\right)(\psi_i^{h,n} - v_i^n) d\boldsymbol{x}.$$

From this, we deduce in a standard fashion that

$$\begin{aligned}\int_{\Omega_t} |\boldsymbol{\psi}^{h,n} - \boldsymbol{v}^n|^2 d\boldsymbol{x} &\leq c_1 \int_{\Omega_0} |\boldsymbol{\psi}^{h,n} - \boldsymbol{v}^n|^2 d\boldsymbol{x} \\ &\quad + c_2 \int_0^T \int_\Omega |\boldsymbol{u}^n - \boldsymbol{u}^{h,n}|^2 \left|\frac{\partial}{\partial x}\boldsymbol{\psi}^{h,n}\right|^2 d\boldsymbol{x}dt,\end{aligned} \tag{3.8}$$

where c_1 and c_2 are independent of h and n.

For a fixed $h > 0$, $\boldsymbol{u}^{h,n} \in L^\infty((0,T);C^\infty(\Omega)^3)$, $\boldsymbol{\psi}^{h,n}(0,\boldsymbol{x}) \in W^{4,2}(\Omega)^3$, and their norms in these spaces are bounded independently of n. From this and the fact that $\boldsymbol{\psi}^{h,n}$ is a solution of a linear partial differential equation with smooth coefficients, we deduce in a standard way that $\boldsymbol{\psi}^{h,n}$ is bounded uniformly in $W^{1,2}((0,T);W^{4,2}(\Omega)^3)$. Therefore, we have that for h fixed and $n \to \infty$ the sequence $\boldsymbol{\psi}^{h,n}$ will converge in $C^0([0,T];W^{4,2}(\Omega)^3)$ to a function $\boldsymbol{\psi}^h$. From this and the weak convergence in L^2 of the sequence $|\boldsymbol{u}^n - \boldsymbol{u}^{h,n}|$ to $|\boldsymbol{w} - \boldsymbol{w}^h|$, it follows that

$$\lim_{n\to\infty} \int_0^T \int_\Omega |\boldsymbol{u}^n - \boldsymbol{u}^{h,n}|^2 \left|\frac{\partial}{\partial x}\boldsymbol{\psi}^{h,n}\right|^2 d\boldsymbol{x}dt \le \int_0^T \int_\Omega |\boldsymbol{w} - \boldsymbol{w}^h|^2 \left|\frac{\partial}{\partial x}\boldsymbol{\psi}^h\right|^2 d\boldsymbol{x}dt,$$

where $\boldsymbol{w}^h$ is the mollification of $\boldsymbol{w}$.

This, together with (3.8), yields in the limit that

$$\begin{aligned}\int_{\Omega_t} |\boldsymbol{\psi}^h - \boldsymbol{w}|^2 d\boldsymbol{x} \le c_1 &\int_{\Omega_0} |\boldsymbol{\psi}^h - \boldsymbol{w}|^2 d\boldsymbol{x}\\ &+ c_2 \int_0^T \int_\Omega |\boldsymbol{w} - \boldsymbol{w}^h|^2 \left|\frac{\partial}{\partial x}\boldsymbol{\psi}^h\right|^2 d\boldsymbol{x}dt.\end{aligned} \tag{3.9}$$

From the definition of $\boldsymbol{\psi}^{h,n}$ and the fact that $\boldsymbol{u}^{h,n}$ converges strongly to $\boldsymbol{w}^h$ as $n \to \infty$, it is easy to see that $\boldsymbol{\psi}^h \equiv \lim_{n\to\infty} \boldsymbol{\psi}^{h,n}$ is the solution to the problem $\frac{\partial \psi_i^h}{\partial t} + w_j^h \frac{\partial \psi_i^h}{\partial x_j} + \frac{\partial p^h}{\partial x_i} = 0$ in $(0,T) \times \Omega$, $i = 1,2,3$, and satisfies $\boldsymbol{\psi}^h(0,\boldsymbol{x}) = \boldsymbol{w}^h(0,\boldsymbol{x})$.

From the uniform regularity of $\boldsymbol{w}^h$, it can be deduced in a standard way that $\boldsymbol{\psi}^h$, the solutions of the associated Oseen equation, will also be uniformly bounded in $W^{1,2}((0,T);W^{3,2}(\Omega)^3)$, for example.

It can then easily be seen that at least for a subsequence of the sequence, $\boldsymbol{\psi}^h$ will converge in $W^{1,2}((0,T);W^{3,2}(\Omega)^3)$ when $h \to 0$. Since the right-hand side of (3.9) goes to zero as $h \to 0$, it follows that, in fact, $\boldsymbol{\psi}^h$ converges to $\boldsymbol{w}$ as $h \to 0$.

Therefore, the sequence $(\boldsymbol{w}^h, \boldsymbol{\psi}^h)$, which is in the graph of the map γ, converges strongly in $C^0([0,T];W^{3,2}(\Omega)^3)$ to $(\boldsymbol{w},\boldsymbol{w})$. Hence $(\boldsymbol{w},\boldsymbol{w})$ is in the strong closure of the graph.

Also, it follows from $\frac{\partial \psi_i^h}{\partial t} + w_j^h \frac{\partial \psi_i^h}{\partial x_j} + \frac{\partial p^h}{\partial x_i} = 0$ and the strong convergence of $(\boldsymbol{w}^h, \boldsymbol{\psi}^h)$ that $\boldsymbol{w}$ is a classical solution of (1.1) in $[0,T) \times \Omega$.

Since T was a generic number less than T_w, it follows that $\boldsymbol{w}$ is a smooth classical solution in $[0,T_w) \times \Omega$. □

Acknowledgments. We would like to thank Professor W. Strauss, whose pointed questions prompted us to write the regularity section, which has strengthened the philosophy of the paper. We would also like to thank Professor V. Sverak for valuable discussions.

REFERENCES

[1] J. Y. Chemin, *Fluides Parfaits Incompressibles*, Astérisque, 230 (1995).

[2] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*, Springer Ser. Comput. Math. 5, Springer-Verlag, Berlin, 1986.

[3] J. L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires*, Dunod Gauthier-Villars, Paris, 1969.

[4] P. L. Lions, *Mathematical Topics in Fluid Mechanics, Vol.* 1: *Incompressible Models*, Oxford Lecture Ser. Math. Appl. 3, Clarendon Press, Oxford, UK, 1996.

[5] J. Málek, J. Nečas, M. Rokyta, and M. Ružička, *Weak and Measure-Valued Solutions to Evolutionary PDE's*, Appl. Math. Math. Comput. 13, Chapman and Hall, London, 1996.

[6] C. Marchioro and M. Pulvirenti, *Mathematical Theory of Incompressible Nonviscous Fluids*, Appl. Math. Sci. 96, Springer-Verlag, New York, 1994.

[7] R. Temam, *Navier-Stokes Equations. Theory and Numerical Analysis*, Stud. Math. Appl. 2, North–Holland, Amsterdam, 1984.

SIAM J. Math. Anal.
Vol. 33, No. 5, pp. 1007–1032

SOME NEW VELOCITY AVERAGING RESULTS*

MICHAEL WESTDICKENBERG†

Abstract. Let (R, μ) be a nonatomic finite measure space, and let $E = L^{\underline{r}}(R, \mu)$ be a Lebesgue space over R. Then we consider tempered distributions f and g (depending on $x \in \mathbb{R}^n$ and $v \in R$), for which $\mathrm{div}_{\mathrm{x}}(\mathbf{a}\mathrm{f}) = \mathrm{g}$ in $\mathcal{S}'(\mathbb{R}^n, E)$. Here $\mathbf{a} : R \longrightarrow \mathbb{R}^n$ is a bounded function of v (a velocity field) satisfying a nondegeneracy condition. We study the regularity of the average $\bar{f} = \int_R f(\cdot, v)\psi(v)\, d\mu(v) \in \mathcal{S}'(\mathbb{R}^n)$ (with $\psi \in L^{\underline{r}'}(R, \mu)$ a suitable weight function) when f and g are bounded in Banach space valued Besov spaces. We also present some compactness results for sequences of averages.

Key words. kinetic equations, velocity averaging, regularity, compactness

AMS subject classifications. 35B65, 35L65, 76P05

PII. S0036141000380760

1. Introduction.

The subject of velocity averaging is the regularity of moments of solutions of transport equations. Let us consider a typical situation: Assume functions f and g are given, depending on $x \in \mathbb{R}^n$ (space) and $v \in \mathbb{R}^n$ (velocity), for which the relation

$$v \cdot \nabla_x f = g \quad \text{in } \mathcal{D}'(\mathbb{R}^n \times \mathbb{R}^n) \tag{1.1}$$

holds. Assume further that we know the regularity of the two functions, e.g., f and g bounded in $L_p(\mathbb{R}^n \times \mathbb{R}^n)$. What can be said about the velocity average

$$\bar{f}(x) = \int_{\mathbb{R}^n} f(x, v)\psi(v)\, dv?$$

Here $\psi \in \mathcal{D}(\mathbb{R}^n)$ is a suitable weight function. It turns out that $\bar{f}$ is somewhat smoother than f and g. Agoshkov [1] showed that if f and g are in $L_2(\mathbb{R}^n \times \mathbb{R}^n)$ and if the weight function is chosen suitably, then the average is bounded in the Sobolev space $W^{1/2,2}(\mathbb{R}^n)$. Hence we have a gain of one half derivative here. Golse, Lions, Perthame, and Sentis [13] proved that $\bar{f} \in W^{1/2,2}(\mathbb{R}^n)$ for all $\psi \in \mathcal{D}(\mathbb{R}^n)$. Their proof is based on a v-dependent decomposition of the Fourier space. Using interpolation, the authors also obtain a result for $1 < p < \infty$: If $f, g \in L_p(\mathbb{R}^n \times \mathbb{R}^n)$, then $\bar{f} \in W^{s,p}(\mathbb{R}^n)$ for all s strictly less than $\min\{1/p, 1/p'\}$.

DiPerna, Lions, and Meyer [10] gave a further improvement. They proved that $\bar{f} \in B^s_{p,\max\{p,2\}}(\mathbb{R}^n)$ with $s = \min\{1/p, 1/p'\}$. Here $B^s_{p,q}(\mathbb{R}^n)$ is a Besov space (cf. section 3.3). Bézard [3] showed that for $1 < p \leq 2$ the average is contained in the (slightly smaller) generalized Sobolev space $H^s_p(\mathbb{R}^n)$. Finally, DeVore and Petrova [7] made clear that $\bar{f} \in B^s_{p,p}(\mathbb{R}^n)$. They also proved that no further improvement with respect to the secondary index q of the Besov norm is possible.

There are several generalizations to the results given. One can assume different integrability for f and g, i.e., $f \in L_p(\mathbb{R}^n \times \mathbb{R}^n)$ and $g \in L_q(\mathbb{R}^n \times \mathbb{R}^n)$ for suitable p, q.

*Received by the editors November 13, 2000; accepted for publication (in revised form) August 22, 2001; published electronically January 16, 2002. This work was supported by Deutsche Forschungsgemeinschaft, SFB 256, at Bonn University.

http://www.siam.org/journals/sima/33-5/38076.html

†Institut für Angewandte Mathematik, Wegelerstrasse 10, D-53115 Bonn, Germany (mwest@iam.uni-bonn.de).

DiPerna, Lions, and Meyer [10] show for this case that the average is contained in a Besov space built on Lorentz spaces instead of $L_p(\mathbb{R}^n)$-spaces as usual. Bézard [3] claims that $\bar{f}$ is even contained in some Sobolev space $H^s_r(\mathbb{R}^n)$, but there is a mistake in his proof.

One can consider also the situation $f, g \in L_q(\mathbb{R}^n_v, L_p(\mathbb{R}^n_x))$ with $1 < q \leq p$. Then it is shown in [10] that $\bar{f} \in B^s_{p,t}(\mathbb{R}^n)$ with $s = \min\{1/q, 1/q'\}$ and suitable t. Note that s depends only on q and that the integrability of $\bar{f}$ is the same as that of f and g, namely, $L_p(\mathbb{R}^n)$. Bézard [3] also studies this situation for $q \leq 2$ and claims that $\bar{f} \in H^s_q(\mathbb{R}^n)$ with $s = 1/p'$; i.e., the roles of p and q are exchanged. But his proof is incorrect.

One can admit derivatives in v or even in x on the right-hand side (RHS) of the transport equation. This was first studied by DiPerna and Lions [9]. But consult also DiPerna, Lions, and Meyer [10]. More general transport operators such as the relativistic streaming operator or transport equations arising from a kinetic formulation of scalar conservation laws can also be considered, cf. Golse, Lions, Perthame, and Sentis [13], DiPerna and Lions [8, 9], Gérard [12], DiPerna, Lions, and Meyer [10], and Lions, Perthame, and Tadmor [16].

The velocity averaging technique can also be used to study compactness. Assume that sequences $f^{(k)}, g^{(k)}$ that satisfy the transport equation (1.1) and are uniformly bounded or precompact in some function space are given; what can be said about the convergence of the sequence $\bar{f}^{(k)}$? It is clear that the regularity results given above imply compactness. However, we refer also to Golse, Perthame, and Sentis [14], Golse, Lions, Perthame, and Sentis [13], DiPerna and Lions [8], Lions, Perthame, and Tadmor [16], Perthame and Souganidis [18], Bouchut [5], and Westdickenberg and Noelle [23].

Finally, let us mention that there is also a relationship between velocity averaging and results known as moments lemmas or dispersion lemmas. Here one considers solutions of the free transport equation $\partial_t f + v \cdot \nabla_x f = 0$ for suitable initial data $f(0, \cdot) = f_0$ with given integrability with respect to x and v. Then one asks what integrability f has in t, x, and v. In particular, one is interested to find decay estimates for f in time. We refer to Perthame [17], Castella and Perthame [6], Bouchut [5], and the references given there.

In this paper, we present some new velocity averaging results. The starting point of our investigation was the question of whether one can gain more than half a derivative in regularity by assuming more integrability in the kinetic variable v. In a sense, our results give an affirmative answer to this question, cf. section 2. We were also interested to find out what the weakest assumptions on f and g are that would still guarantee strong precompactness. We give some answers to that question below.

This paper is organized as follows: In section 2 we first develop a new view on velocity averaging and state our regularity and compactness results. In section 3 we then collect some facts from the theory of Banach space valued tempered distributions. Sections 4 and 5 contain the proofs of our results.

We will assume in the following the space dimension $n \geq 2$. We denote by $\mathcal{D}(\mathbb{R}^n)$ the space of C^∞-functions with compact support, equipped with the usual topology of test functions. $\mathcal{D}'(\mathbb{R}^n)$ is the corresponding dual, the space of distributions.

2. Main results. We want to make an attempt here to develop a somewhat different, less pragmatic view on velocity averaging than is usual in the literature. Therefore, we go back to the transport equation $\operatorname{div}_x(\mathbf{a}f) = g$ and make precise in what sense this equation should hold. We will use notions and results from the theory of Banach space valued tempered distributions. The reader is referred to section 3, where we put together some information relevant to our discussion.

Let (R,μ) be a nonatomic finite measure space, and let $E = L_{\underline{r}}(R,\mu)$, $1 \le \underline{r} \le \infty$, be a Lebesgue space over (R,μ). Then we consider distributions f and g in $\mathcal{S}'(\mathbb{R}^n, E)$, i.e., linear mappings of the Schwartz class $\mathcal{S}(\mathbb{R}^n)$ (x-dependance) into E (v-dependance), that are continuous with respect to the Fréchet topology of $\mathcal{S}(\mathbb{R}^n)$. We assume that a function $\mathbf{a}\colon R \longrightarrow \mathbb{R}^n$ (velocity field) in $L_\infty(R,\mu)$ is given and that the following relation holds for f and g:

$$\operatorname{div}_x(\mathbf{a}f) = g \qquad \text{in } \mathcal{S}'(\mathbb{R}^n, E). \tag{2.1}$$

By the definition of $\mathcal{S}'(\mathbb{R}^n, E)$, this means that for all test functions $\varphi \in \mathcal{S}(\mathbb{R}^n)$

$$-\sum_{j=1}^{n} \mathbf{a}_j \langle f, \partial_j \varphi\rangle = \langle g, \varphi\rangle \qquad \text{in } E, \tag{2.2}$$

i.e., μ-almost everywhere (μ-a.e.). We use brackets to denote the dual pairing of distributions and test functions. Note that multiplication with $\mathbf{a} \in L_\infty(R,\mu)$ maps E continuously into itself. If both f and g are regular, we may also write

$$-\sum_{j=1}^{n} \mathbf{a}_j \int_{\mathbb{R}^n} f(x,\cdot)\partial_j\varphi(x)\,dx = \int_{\mathbb{R}^n} g(x,\cdot)\varphi(x)\,dx \qquad \text{in } E.$$

Now let ψ be an element of the conjugate space $E' = L_{\underline{r}'}(R,\mu)$ with $1/\underline{r} + 1/\underline{r}' = 1$. Then we can define the average $\bar{f}$ to be that distribution in $\mathcal{S}'(\mathbb{R}^n)$ for which

$$\langle \bar{f}, \varphi\rangle = \int_R \langle f(\cdot, v), \varphi\rangle\, \psi(v)\, d\mu(v) \quad \forall \varphi \in \mathcal{S}(\mathbb{R}^n). \tag{2.3}$$

Hence the average is the pairing of $f \in \mathcal{S}'(\mathbb{R}^n, E)$ with some $\psi \in E'$. Assume now that boundedness of f and g in suitable function spaces is given. Then we may ask for the regularity of $\bar{f}$.

2.1. Regularity. We will use Banach space valued Besov spaces $B^s_{p,q}(\mathbb{R}^n, E)$ with $s \in \mathbb{R}$ and $0 < p,\, q \le \infty$. Assume that we are given two tempered distributions

$$f \in B^s_{p,q}(\mathbb{R}^n, E_1) \quad \text{and} \quad g \in B^{s-\tau}_{p,q}(\mathbb{R}^n, E_2)$$

for suitable numbers $s, \tau \in \mathbb{R}$ and spaces $E_1 = L_{r_1}(R,\mu)$ and $E_2 = L_{r_2}(R,\mu)$. We will consider two different cases: $0 < p \le 1$ (Case I) and $1 < p < \frac{n}{n-1}$ (Case II). We put $E = L_{\underline{r}}(R,\mu)$ with $\underline{r} = \min\{r_1, r_2\}$ and assume that (2.1) holds.

Now let $F = L_r(R,\mu)$ be another Lebesgue space with $1/r \le \min\{1/r_1', 1/r_2'\}$. Then multiplication by some $\psi \in F$ maps E_1 continuously into $L_{\rho_1}(R,\mu)$, where $1/\rho_1 = 1/r + 1/r_1$, and analogously maps E_2 into some $L_{\rho_2}(R,\mu)$. We fix a weight $\psi \in F$ and define the average $\bar{f}$ as in (2.3).

We will assume further that the velocity field $\mathbf{a}$ is nondegenerate in the following sense: There are numbers $C > 0$ and $\alpha \in (0,1]$ s.t. for all $\delta \in (0,1]$

$$\sup_{\xi \in \mathbb{R}^n} \mu\{v \in R \colon |\mathbf{a}(v)\cdot \xi/|\xi|| \le \delta\} \le C\delta^\alpha. \tag{2.4}$$

Remark 2.1. This condition was first used in [16]. Let us give an example. If R is a compact subset of $\mathbb{R}^n$ and μ is the Lebesgue measure, then, for $\mathbf{a}(v) = v$, (2.4) is satisfied with $\alpha = 1$. This velocity field appears, e.g., in the Boltzmann equations.

THEOREM 2.2. *Let $0 < q \le \infty$, $0 < \tau \le 1$, and $\alpha/\rho_2' < 1$. With the assumptions above, $\bar{f}$ is bounded in $B^S_{P,q}(\mathbb{R}^n)$ for numbers P and $S = s - \kappa + \Delta S$ given by*

Case I	$0 < p \le 1$	$P = 2$	$\kappa = n\left(\frac{1}{p} - \frac{1}{2}\right)$
Case II	$1 < p < \frac{n}{n-1}$	$P = \left[\frac{1}{p} - \frac{n-1}{n}\right]^{-1}$	$\kappa = n - 1$

and

$$\Delta S = (1-\tau)\frac{\alpha}{\rho_1'}\left[1 + \alpha\left(\frac{1}{\rho_2} - \frac{1}{\rho_1}\right)\right]^{-1}. \tag{2.5}$$

More precisely, there exists some constant $C > 0$ s.t. for all f, g, and ψ in the respective function spaces the following inequality holds:

$$\|\bar{f}\|_{B^S_{P,q}(\mathbb{R}^n)} \le C\|\psi\|_F \left\{\|f\|_{B^s_{p,q}(\mathbb{R}^n,E_1)} + \|g\|_{B^{s-\tau}_{p,q}(\mathbb{R}^n,E_2)}\right\} \tag{2.6}$$

whenever f and g satisfy the transport equation (2.1).

Remark 2.3. The regularity of the average differs from that of f (i.e., s) by two terms. First, we lose $n(\frac{1}{p} - \frac{1}{P})$ derivatives. More precisely, we change regularity for integrability: Instead of $L_p(\mathbb{R}^n)$-boundedness we now have $L_P(\mathbb{R}^n)$. This is simply a Sobolev imbedding. Second, we gain some regularity ΔS, which is a nonnegative number. This regularizing effect is an outcome of the nondegeneracy of $\mathbf{a}$. Note that ΔS does not depend on p and q. It is a function of α, τ, ρ_1, and ρ_2 only.

Let us discuss a few special cases. If $\alpha = 1$, $\tau = 0$, and $r = \infty$, then $\Delta S \to 1$ for r_1 and r_2 getting large. In other words, we gain almost a full derivative if we have much integrability with respect to the kinetic variable v. (Remember that (R, μ) is a finite measure space.) For $r_1 = r_2 = 2$ we find $\Delta S = 1/2$. Now suppose that g is a full derivative less regular than f, i.e., $\tau = 1$; then $\Delta S = 0$. This is obvious since in that case the transport equation contains no nontrivial regularity information. Note also that ΔS is getting smaller if r is chosen small: For more general (less integrable) weights $\psi \in L_r(R, \mu)$ we pay with a smaller gain of regularity.

Here is an example: If we consider the transport equation $v \cdot \nabla_x f = g$, for which the nondegeneracy condition (2.4) holds with $\alpha = 1$, and if both f and g are contained in $B^0_{1,1}(\mathbb{R}^n, L_\infty(R, \mu))$, then for $\psi \in L_\infty(R, \mu)$ the average $\bar{f}$ is in the negative Sobolev space $H^\epsilon_2(\mathbb{R}^n)$ for all $\epsilon < -\frac{n}{2} + 1$. In the particular case $n = 2$, this means that the average is almost in $L_2(\mathbb{R}^n)$. We cannot reach $\epsilon = 0$ because α/ρ_2' has to be strictly less than 1 (i.e., although the velocity field $\mathbf{a}$ would admit $\alpha = 1$, in our estimates we have to use an $\alpha < 1$). This is an improvement over the results mentioned in section 1 since in section 1 at most quadratic integrability in v could be handled, giving a maximal gain of regularity of one half derivative. Note that for an arbitrary Banach space E the following embedding holds:

$$B^0_{1,1}(\mathbb{R}^n, E) \hookrightarrow L_1(\mathbb{R}^n, E) \hookrightarrow B^0_{1,\infty}(\mathbb{R}^n, E). \tag{2.7}$$

(This follows immediately from the definitions.) Therefore, if we start with f and g in $L_1(\mathbb{R}^n, L_\infty(R, \mu))$, we obtain $\bar{f} \in B^\epsilon_{2,\infty}(\mathbb{R}^n)$ with ϵ as above. The latter space is slightly larger than the corresponding Sobolev space $H^\epsilon_2(\mathbb{R}^n)$.

Remark 2.4. Clearly, it would be nice to get rid of the Sobolev embedding at least partially. And it is also a little bit disappointing that the two cases do not match for $p = 1$ unless $n = 2$: The P for $p > 1$ is considerably larger than that for $p \le 1$. (Hence we lose more derivatives in the Sobolev embedding.) The reason for this discrepancy is that we employ different methods of proof for the two cases of our theorem: for the first case we use a decomposition of f and g into simple building blocks, which gives a decoupling of the x- and v-dependance; for the second case we use the Radon transform and its regularizing properties (cf. section 4). We are not aware of a straightforward way to connect these two methods and thereby fill the gap in P mentioned above. Nevertheless, we hope that our methods are interesting for their own sake.

2.2. Compactness. We now want to discuss a few compactness results and start with a generalization of what was said in section 2.1. Let (R, μ) again be a nonatomic finite measure space, and let E_1 and E_2 be two arbitrary rearrangement-invariant Banach function spaces (cf. section 5.1). Then we consider sequences of distributions bounded in Banach space valued Besov spaces,

$$f^{(k)} \in B^s_{p,q}(\mathbb{R}^n, E_1) \quad \text{and} \quad g^{(k)} \in B^{s-\tau}_{p,q}(\mathbb{R}^n, E_2), \tag{2.8}$$

that satisfy the transport equation (2.1) in $\mathcal{S}'(\mathbb{R}^n, E)$ with $E = E_1 + E_2$. Assume there exists a subset F of the associated space E' of E and two further Banach function spaces G_1 and G_2 such that multiplication with $\psi \in F$ maps E_1 continuously into G_1, and E_2 into G_2. We are interested in the precompactness of the sequence of averages $\bar{f}^{(k)}$ (cf. definition (2.3)) in local Besov spaces $B^{S,\text{loc}}_{P,q}(\mathbb{R}^n)$.

More precisely, we want to identify circumstances under which the sequence of products $\chi \bar{f}^{(k)}$ contains a subsequence converging in $B^S_{P,q}(\mathbb{R}^n)$, where $\chi \in \mathcal{D}(\mathbb{R}^n)$ is an arbitrary test function with compact support. Again we must assume nondegeneracy of the velocity field $\mathbf{a}$, which now takes the form

$$\lim_{\delta \to 0} \eta(\delta) = 0, \qquad \text{where } \eta(\delta) = \sup_{\xi \in \mathbb{R}^n} \mu \left\{ v \in R \colon |\mathbf{a}(v) \cdot \xi / |\xi|| \le \delta \right\}. \tag{2.9}$$

Note that this assumption is weaker than condition (2.4). Again we consider two different situations: $0 < p \le 1$ and $1 < p < \frac{n}{n-1}$. Then we have the following theorem.

THEOREM 2.5. *Let $0 < q \le \infty$ and $0 < \tau < 1$. Assume that the fundamental function of the associated space G_1' of G_1 is continuous at zero. Then the sequence of averages $\bar{f}^{(k)}$ is precompact in $B^{S,\text{loc}}_{P,q}(\mathbb{R}^n)$, where $S = s_1 - \kappa$ with the following.*

Case I	$0 < p \le 1$	$P = 2$	$\kappa = n \left(\frac{1}{p} - \frac{1}{2} \right)$
Case II	$1 < p < \frac{n}{n-1}$	$P = \left[\frac{1}{p} - \frac{n-1}{n} \right]^{-1}$	$\kappa = n - 1$

If $\tau = 1$, we still have precompactness if we assume that the sequence $g^{(k)}$ is not only bounded in $B^{s-\tau}_{p,q}(\mathbb{R}^n, E_2)$ but strongly precompact.

For the definition of the fundamental function we refer to section 5.1.

Remark 2.6. The gain of regularity due to the averaging process depends primarily on the nondegeneracy of the velocity field $\mathbf{a}$ and the integrability of f with respect to the kinetic variable v: the higher α and ρ_1 are in Theorem 2.2, the bigger ΔS becomes. Here we consider a situation where $\mathbf{a}$ is only weakly nondegenerate (i.e., we assume only that $\eta(\delta) \to 0$ for $\delta \to 0$; nothing is said about a polynomial rate) and where also the v-integrability of $f^{(k)}$ is only slightly better than $L_1(R, \mu)$. Then we still have precompactness, but there is no gain of regularity at all. Of course, if we strengthen our assumptions, we will get more: If instead of (2.9) we have (2.4) and if E_1, E_2 are Lebesgue spaces as above, then we can combine the proofs of Theorems 2.2 and 2.5 to show precompactness in $B^{S,\mathrm{loc}}_{P,q}(\mathbb{R}^n)$ with regularity S strictly less than $s - n(\frac{1}{p} - \frac{1}{P}) + \Delta S$ and ΔS given by (2.5).

Remark 2.7. Let us discuss a few examples. The choice $E_1 = L^1(R, \mu)$ forces $F = L^\infty(R, \mu)$ and hence $G_1 = L^1(R, \mu)$ because $L^1(R, \mu)$ is the largest of all rearrangement-invariant Banach function spaces over (R, μ). The fundamental function of the corresponding associated space $G'_1 = L^\infty(R, \mu)$ is discontinuous at zero (cf. section 5.1). Therefore, Theorem 2.5 does not apply: You need more than simple L^1-integrability in the kinetic variable v to obtain strong compactness.

A sufficient condition would be, for example, $E_1 = L \log L(R, \mu)$ and $F = L^\infty(R, \mu)$. Then $G_1 = L \log L(R, \mu)$, and the fundamental function of $G'_1 = \exp L(R, \mu)$ is continuous at zero. $L \log L$-integrability plays an important role for the Boltzmann equations: If f is a solution of this system, then $f \log f$ is the entropy density, and the famous H-theorem tells us that the global entropy does not increase in time.

Note, finally, that the choice $F = E'_1$ does not work either, since from the Hölder inequality (5.1) we again obtain only L^1-integrability for the product.

Concretely, if $f^{(k)}$ and $g^{(k)}$ are bounded in the Besov spaces $B^0_{1,1}(\mathbb{R}^n, L \log L(R, \mu))$ and $B^{-\tau}_{1,1}(\mathbb{R}^n, L_1(R, \mu))$ with $\tau < 1$, then for $\psi \in L_\infty(R, \mu)$ the average $\bar{f}^{(k)}$ is locally precompact in the Sobolev space $H^\epsilon_2(\mathbb{R}^n)$ with $\epsilon = -\frac{n}{2}$. We can admit $\tau = 1$ if we assume strong precompactness instead of mere boundedness for $g^{(k)}$. If $f^{(k)}$ is only bounded in $L_1(\mathbb{R}^n, L \log L(R, \mu))$ instead, then we can use the embedding (2.7) to obtain local precompactness of $\bar{f}^{(k)}$ in $B^\epsilon_{2,\infty}(\mathbb{R}^n)$ with ϵ as above.

Remark 2.8. To prove precompactness of $\bar{f}^{(k)}$ it is sufficient to have boundedness of $f^{(k)}$ and $g^{(k)}$ in local Besov spaces only, cf. Remark 5.9.

3. Preliminaries.

We collect here some results we will need later in the proofs. We start with a few remarks about Banach space valued distributions.

3.1. Banach space valued tempered distributions.

If E is some arbitrary Banach space, we define the Schwartz class $\mathcal{S}(\mathbb{R}^n, E)$ to be the space of infinitely differentiable, rapidly decreasing functions on $\mathbb{R}^n$ taking their values in E. This space is locally convex and complete with respect to the Fréchet topology defined by the family of seminorms

$$p_N(\varphi) = \sup_{|\alpha| \le N} \sup_{x \in \mathbb{R}^n} (1 + |x|)^N \|\partial^\alpha \varphi(x)\|_E \quad \text{with } N \in \mathbb{N}_0.$$

We abbreviate $\mathcal{S}(\mathbb{R}^n) = \mathcal{S}(\mathbb{R}^n, \mathbb{C})$. Now let $\mathcal{O}_M(\mathbb{R}^n, E)$ be the space of smooth E-valued functions with at most polynomial growth at infinity (also called slowly increasing). Again the topology is defined by a family of seminorms

$$\psi \longmapsto \sup_{x \in \mathbb{R}^n} \|\varphi(x) \partial^\alpha \psi(x)\|_E \quad \text{for } \varphi \in \mathcal{S}(\mathbb{R}^n), \alpha \in \mathbb{N}^n_0. \tag{3.1}$$

We denote by $\mathcal{S}'(\mathbb{R}^n, E)$ the space of linear mappings from $\mathcal{S}(\mathbb{R}^n)$ into E that are continuous with respect to the strong topology of the Schwartz class. The dual pairing of some $f \in \mathcal{S}'(\mathbb{R}^n, E)$ with a test function $\varphi \in \mathcal{S}(\mathbb{R}^n)$ is expressed using brackets: $\langle f, \varphi \rangle$. Note that this quantity is an element of E. Therefore, equality in $\mathcal{S}'(\mathbb{R}^n, E)$ means equality in E after testing against $\varphi \in \mathcal{S}(\mathbb{R}^n)$. If $f \in L^1_{\mathrm{loc}}(\mathbb{R}^n, E)$, the pairing is just an integral. As we did for $\mathcal{S}(\mathbb{R}^n)$, we will simply write $\mathcal{O}_M(\mathbb{R}^n)$ and $\mathcal{S}'(\mathbb{R}^n)$ whenever $E = \mathbb{C}$.

Exactly as in the scalar case, we can define derivatives of E-valued tempered distributions $f \in \mathcal{S}'(\mathbb{R}^n, E)$ or the product of f with a slowly increasing function in $\mathcal{O}_M(\mathbb{R}^n)$. Additionally, the notions of support $\operatorname{supp} f$ of f, Fourier transform $\hat{f} = \mathbf{F}f$ and its inverse $\check{f} = \mathbf{F}^{-1}f$, and convolution $f \star \rho$ for $\rho \in \mathcal{S}(\mathbb{R}^n)$ can be carried over from the scalar theory. Again we have the identity $f \star \rho = \mathbf{F}^{-1}[\hat{\rho}\mathbf{F}f]$ in $\mathcal{S}'(\mathbb{R}^n, E)$. Instead of going into details here, we refer to Amann [2] or Hörmander [15].

Let us assume now that besides E there exist two more Banach spaces F and G and a bilinear continuous mapping $\cdot \colon F \times E \longrightarrow G$ with a norm not bigger than one. We call this mapping a multiplication. Then we can define the product $a \bullet f$ of some E-valued tempered distribution $f \in \mathcal{S}'(\mathbb{R}^n, E)$ with a function $a = \psi \otimes \chi$, where $\psi \in \mathcal{O}_M(\mathbb{R}^n)$ and $\chi \in F$, to be that distribution in $\mathcal{S}'(\mathbb{R}^n, G)$ for which

$$\langle a \bullet f, \varphi \rangle = \chi \cdot \langle f, \psi\varphi \rangle \quad \forall \varphi \in \mathcal{S}(\mathbb{R}^n). \tag{3.2}$$

Note that $\langle f, \psi\varphi \rangle \in E$. Then we can ask whether that definition can be extended to a class of functions more general than $a = \psi \otimes \chi$.

Theorem 3.1. *There exists a uniquely defined bilinear mapping*

$$\begin{aligned} \bullet \colon \mathcal{O}_M(\mathbb{R}^n, F) \times \mathcal{S}'(\mathbb{R}^n, E) &\longrightarrow \mathcal{S}'(\mathbb{R}^n, G), \\ (a, f) &\longmapsto a \bullet f, \end{aligned}$$

with (3.2) *for all $a = \psi \otimes \chi$. The mapping $\bullet$ is uniformly continuous with respect to each variable if the respective other one is confined to bounded subsets.*

Remark 3.2. This is a special case of the Schwartz kernel theorem (cf. Theorem 2.1 in Amann [2]). The idea of the proof is the following: Choose some test function $\eta \in \mathcal{D}(\mathbb{R}^n)$ with $\eta \geq 0$ and $\int_{\mathbb{R}^n} \eta(x)\, dx = 1$. For $\epsilon > 0$ define

$$\eta_\epsilon(x) = \epsilon^{-n}\eta(x/\epsilon) \quad \forall x \in \mathbb{R}^n.$$

Then the convolution $f_\epsilon = f \star \eta_\epsilon$ is a function in $\mathcal{O}_M(\mathbb{R}^n, E)$ with $\lim_{\epsilon \to 0} f_\epsilon = f$ in $\mathcal{S}'(\mathbb{R}^n, E)$. If $a \in \mathcal{O}_M(\mathbb{R}^n, F)$ is given, we can define the product $a \bullet f_\epsilon$ pointwise by $a \bullet f_\epsilon(x) = a(x) \cdot f_\epsilon(x) \in G$ for all $x \in \mathbb{R}^n$. Then we put

$$a \bullet f = \lim_{\epsilon \to 0} a \bullet f_\epsilon \quad \text{in } \mathcal{S}'(\mathbb{R}^n, G).$$

Of course, a little work is necessary to show that this definition makes sense. If there is no danger of confusion, we will abbreviate $a \bullet f$ by af.

Remark 3.3. We also need the following fact. Assume there exist Banach spaces E, F_1, F_2, G, H_1, H_2 and multiplications (all denoted by $\cdot$)

$$\begin{tikzcd}
F_1 & \times \arrow[d] & F_2 & & & & F_2 & \times \arrow[d] & E \\
& H_1 & \times & E \arrow[r] & G & F_1 \arrow[l] & \times & H_2 &
\end{tikzcd}$$

that are associative; i.e., $(\chi_1 \cdot \chi_2) \cdot e = \chi_1 \cdot (\chi_2 \cdot e)$ in G for all $\chi_j \in F_j$ and $e \in E$. Then the pointwise multiplication of Theorem 3.1 is associative as well. More precisely,

$$(M_1 \bullet M_2) \bullet f = M_1 \bullet (M_2 \bullet f) \quad \text{in } \mathcal{S}'(\mathbb{R}^n, G)$$

for all $M_j \in \mathcal{O}_M(\mathbb{R}, F_j)$ and $f \in \mathcal{S}'(\mathbb{R}^n, E)$ (cf. Amann [2]).

3.2. L_p-spaces of Banach space valued entire functions. Let $f \in \mathcal{S}'(\mathbb{R}^n, E)$ be a distribution with the property that the support of its Fourier transform $\mathbf{F}f$ is contained in a compact subset $K \subset \mathbb{R}^n$. If $\rho \in \mathcal{S}(\mathbb{R}^n)$ with $\operatorname{supp} \hat{\rho}$ compact and $\hat{\rho}(\xi) = 1$ for all $\xi \in K$, then the identity $f = \mathbf{F}^{-1}[\hat{\rho}\mathbf{F}f] = f \star \rho$ in $\mathcal{S}'(\mathbb{R}^n, E)$ follows immediately from the definitions. Now ρ is an entire analytic function on $\mathbb{R}^n$ that can be extended to $\mathbb{C}^n$. This is the famous Payley–Wiener–Schwartz theorem. For any $N \in \mathbb{N}$ there exists some constant $C_N > 0$ s.t.

$$|\rho(z)| \le C_N(1+|z|)^{-N} e^{c|\mathrm{Im} z|} \quad \forall z \in \mathbb{C}^n. \tag{3.3}$$

From this estimate one derives, completely analogous to the scalar valued case, that $f \in \mathcal{S}'(\mathbb{R}^n, E)$ with $\operatorname{supp} \mathbf{F}f$ compact is an entire analytic E-valued function too. We refer to [15, Theorem 7.3.1] for the argument with $E = \mathbb{C}$.

Definition 3.4. *Let E be some Banach space with $K \subset \mathbb{R}^n$ compact and $0 < p \le \infty$. Then we define the L_p-space of E-valued entire analytic functions*

$$L_{p,K}(\mathbb{R}^n, E) = \Big\{ f \in \mathcal{S}'(\mathbb{R}^n, E) \colon \operatorname{supp} \hat{f} \subset K, \|f\|_{L_p(\mathbb{R}^n,E)} < \infty \Big\}$$

$$\textit{with} \quad \|f\|_{L_p(\mathbb{R}^n,E)} = \left(\int_{\mathbb{R}^n} \|f(x)\|_E^p \, dx \right)^{1/p}.$$

The space $L_{p,K}(\mathbb{R}^n, E)$ is complete with respect to $\|\cdot\|_{L_p(\mathbb{R}^n,E)}$.

Theorem 3.5. *Let E be some Banach space with $K \subset \mathbb{R}^n$ compact, $0 < p \le \infty$, and $0 < w < \infty$. Then there exists a constant $C > 0$ s.t. for all $f \in L_{p,K}(\mathbb{R}^n, E)$*

$$\sup_{y \in \mathbb{R}^n} \|f(x-y)\|_E \, (1+|y|)^{-n/w} \le C \left(\mathbf{M}\|f\|_E^w \right)^{1/w}(x) \quad \forall x \in \mathbb{R}^n.$$

Here $\mathbf{M}$ is the usual Hardy–Littlewood maximal operator

$$\mathbf{M}g(x) = \sup \left\{ \frac{1}{|Q|} \int_Q |g(y)| \, dy \colon \textit{all cubes } Q \textit{ containing } x \right\}.$$

Now if $0 < w < p$ (hence $p/w > 1$), we obtain as an immediate consequence of the Hardy–Littlewood maximal inequality (cf. Stein [20])

$$\begin{aligned} \left\| \sup_{y \in \mathbb{R}^n} \frac{\|f(\cdot - y)\|_E}{(1+|y|)^{n/w}} \right\|_{L^p(\mathbb{R}^n)} &\le C \, \|\mathbf{M}\|f\|_E^w\|_{L^{p/w}(\mathbb{R}^n)}^{1/w} \\ &\le C \, \|\|f\|_E^w\|_{L^{p/w}(\mathbb{R}^n)}^{1/w} = C\|f\|_{L_p(\mathbb{R}^n,E)}. \end{aligned}$$

The constant $C = C(n, p, K, w)$ does not depend on f.

Theorem 3.6 (Nikol'skij inequality). *Let E and K be as above with $0 < p \le q \le \infty$ and $\alpha \in \mathbb{N}_0^n$. Then there exists a constant $C > 0$ s.t. for all $f \in L_{p,K}(\mathbb{R}^n, E)$*

$$\|\partial^\alpha f\|_{L_q(\mathbb{R}^n,E)} \le C\|f\|_{L_p(\mathbb{R}^n,E)}. \tag{3.4}$$

We refer to Triebel [22, Chapter III/15] and to the literature cited there.

3.3. Banach space valued Besov spaces. Let φ_0 be a radially symmetric test function in $\mathcal{S}(\mathbb{R}^n)$ supported in $B_2(0) \subset \mathbb{R}^n$ with $\varphi_0(\xi) = 1$ for all $|\xi| \le 1$. Then define $\varphi_1(\xi) = \varphi_0(2^{-1}\xi) - \varphi_0(\xi)$ and $\varphi_\nu(\xi) = \varphi_1(2^{-\nu+1}\xi)$ for $\nu \in \mathbb{N}$. We obtain a dyadic decomposition of unity $\sum_{\nu \in \mathbb{N}_0} \varphi_\nu(\xi) = 1$ for all $\xi \in \mathbb{R}^n$.

DEFINITION 3.7. *Let E be an arbitrary Banach space with $0 < p,\, q \leq \infty$, and $s \in \mathbb{R}$. Then the Banach space valued Besov space $B^s_{p,q}(\mathbb{R}^n, E)$ is defined as the space of all E-valued distributions $f \in \mathcal{S}'(\mathbb{R}^n, E)$ for which the Besov (quasi) norm*

$$\|f\|_{B^s_{p,q}(\mathbb{R}^n,E)} = \Big\| \big\{ 2^{\nu s} \|f_\nu\|_{L_p(\mathbb{R}^n,E)} \big\}_\nu \Big\|_{\ell_q(\mathbb{N}_0)} \tag{3.5}$$

is finite. Here $f_\nu = \mathbf{F}^{-1}[\varphi_\nu \hat{f}]$ in $\mathcal{S}'(\mathbb{R}^n, E)$. For $1 \leq p, q \leq \infty$, (3.5) is a norm.

Remark 3.8. We stress that the Banach space is completely arbitrary. Assumptions like the unconditionality of Martingale differences (UMD) property, separability, or reflexivity are not necessary. If $E = \mathbb{C}$, we will simply write $B^s_{p,q}(\mathbb{R}^n)$.

Remark 3.9. Assume $t \in \mathbb{R}$, $0 < r \leq \infty$, and let F be a second Banach space continuously embedded into E. As an immediate consequence of definition (3.5), the following inclusions hold:

$$\begin{aligned} &B^s_{p,q}(\mathbb{R}^n, E) \hookrightarrow B^s_{p,r}(\mathbb{R}^n, E) \quad \text{if } q \leq r,\\ &B^s_{p,q}(\mathbb{R}^n, E) \hookrightarrow B^t_{p,q}(\mathbb{R}^n, E) \quad \text{if } t \leq s, \text{ and}\\ &B^s_{p,q}(\mathbb{R}^n, E) \hookrightarrow B^s_{p,q}(\mathbb{R}^n, F). \end{aligned}$$

For any $\sigma \in \mathbb{R}$ the operator $\mathcal{J}_\sigma$, defined by

$$\widehat{\mathcal{J}_\sigma \varphi}(\xi) = \left(1 + |\xi|^2\right)^{\sigma/2} \hat{\varphi}(\xi) \quad \forall \xi \in \mathbb{R}^n,$$

maps the Schwartz class $\mathcal{S}(\mathbb{R}^n)$ injectively onto itself. A posteriori then, the same is true for the space $\mathcal{S}'(\mathbb{R}^n, E)$ because the product of $\mathbf{F}f \in \mathcal{S}'(\mathbb{R}^n, E)$ with a function $(1 + |\cdot|^2)^{\sigma/2} \in \mathcal{O}_M(\mathbb{R}^n)$ is well defined. We have $\mathcal{J}_\sigma \circ \mathcal{J}_{-\sigma} = \mathrm{Id}$ in $\mathcal{S}'(\mathbb{R}^n, E)$.

THEOREM 3.10. *For numbers $s \in \mathbb{R}$ and $0 < p, q \leq \infty$, the two spaces $B^{pq}_{s-\sigma}(\mathbb{R}^n, E)$ and $\mathcal{J}_\sigma B^s_{p,q}(\mathbb{R}^n, E) = \big\{ \mathcal{J}_\sigma f \colon f \in B^s_{p,q}(\mathbb{R}^n, E) \big\}$ coincide. The quantity*

$$\|f\|^*_{B^s_{p,q}(\mathbb{R}^n,E)} = \|\mathcal{J}_\sigma f\|_{B^{s-\sigma}_{p,q}(\mathbb{R}^n,E)}$$

is an equivalent (quasi) norm on $B^s_{p,q}(\mathbb{R}^n, E)$.

Remark 3.11. So Besov spaces of different regularity (with the same p, q, of course) are isomorphic to each other. The mapping $\mathcal{J}_\sigma$ is called a lifting. Let us also recall the closely related estimate

$$\|\partial^\alpha f\|_{B^s_{p,q}(\mathbb{R}^n,E)} \leq C \|f\|_{B^{s+|\alpha|}_{p,q}(\mathbb{R}^n,E)} \quad \forall \alpha \in \mathbb{N}^n_0.$$

We will not give here the proof of Theorem 3.10 nor that of the next one, Theorem 3.12. In both cases, it is an easy adaptation of the corresponding result for the scalar case. We refer to Triebel [21].

THEOREM 3.12. *Let $s \in \mathbb{R}$, $0 < p, q \leq \infty$, and $\psi \in \mathcal{O}_M(\mathbb{R}^n)$. Then for large M*

$$\|\psi f\|_{B^s_{p,q}(\mathbb{R}^n,E)} \leq C \sum_{|\alpha| \leq M} \|\partial^\alpha \psi\|_{L_\infty(\mathbb{R}^n)} \|f\|_{B^s_{p,q}(\mathbb{R}^n,E)} \quad \forall f \in \mathcal{S}'(\mathbb{R}^n, E). \tag{3.6}$$

4. Proofs—regularity. We briefly repeat the assumptions made in section 2.1. Let (R, μ) be a nonatomic finite measure space, and fix Lebesgue spaces $E_1 = L_{r_1}(R, \mu)$ and $E_2 = L_{r_2}(R, \mu)$ with $1 \leq r_1, r_2 \leq \infty$. Assume we are given $f \in B^s_{p,q}(\mathbb{R}^n, E_1)$ and $g \in B^{s-\tau}_{p,q}(\mathbb{R}^n, E_2)$ for some $0 < p, q \leq \infty$ and $s, \tau \in \mathbb{R}$ satisfying a transport equation $\mathrm{div}_x(\mathbf{a}f) = g$ in $\mathcal{S}'(\mathbb{R}^n, E)$, where $E = L_{\underline{r}}(R, \mu)$ and

$\underline{r} = \min\{r_1, r_2\}$. For an arbitrary $\psi \in F = L_r(R,\mu)$ with $1/r \leq \min\{1/r_1', 1/r_2'\}$ define the average $\bar{f} \in \mathcal{S}'(\mathbb{R}^n)$ by

$$\langle \bar{f}, \phi \rangle = \int_R \langle f(\cdot, v), \phi \rangle \, \psi(v) \, d\mu(v) \quad \forall \phi \in \mathcal{S}(\mathbb{R}^n).$$

We will prove that if the velocity field $\mathbf{a}$ satisfies a nondegeneracy condition (2.4), and provided that suitable assumptions on the various parameters hold (which we will repeat in the course of the proof as they are needed), then the average $\bar{f}$ is bounded in the Besov space $\bar{f} \in B^S_{P,q}(\mathbb{R}^n)$ for numbers P, S given in section 2.1.

Without restriction of generality we may fix some s large because of a simple lifting argument: The operator $\mathcal{J}_\sigma$ introduced in section 3.3 is an isomorphism between Banach space valued Besov spaces, $\mathcal{J}_\sigma B^s_{p,q}(\mathbb{R}^n, E) \approx B^{s-\sigma}_{p,q}(\mathbb{R}^n, E)$, and it commutes with the transport operator. We may therefore consider the distributions $F = \mathcal{J}_\sigma f$, respectively, $G = \mathcal{J}_\sigma g$. They satisfy the same transport equation $\operatorname{div}_x(\mathbf{a}F) = G$. Then $\bar{F}$ is equal to $\mathcal{J}_\sigma \bar{f}$, which means that the regularity is simply shifted by σ.

The average $\bar{f}$ is a tempered distribution and can therefore be decomposed into its dyadic components $\bar{f} = \sum_{\nu=0}^\infty \bar{f}_\nu$ in $\mathcal{S}'(\mathbb{R}^n)$ with $\bar{f}_\nu = \mathbf{F}^{-1}[\varphi_\nu \mathbf{F} \bar{f}]$ as usual. It is then sufficient to estimate each block separately: We will show that there exists a constant $C > 0$ independent of f and g s.t. for all $\nu \geq 0$

$$(4.1) \qquad 2^{\nu S} \|\bar{f}_\nu\|_{L_P(\mathbb{R}^n)} \leq C \|\psi\|_F \left\{ 2^{\nu s} \|f_\nu\|_{L_p(\mathbb{R}^n, E_1)} + 2^{\nu(s-\tau)} \|g_\nu\|_{L_p(\mathbb{R}^n, E_2)} \right\}.$$

Here f_ν and g_ν are the dyadic components of f and g. We take the $\ell_q(\mathbb{N}_0)$-norm of the sequence $\left\{2^{\nu S} \|\bar{f}_\nu\|_{L_P(\mathbb{R}^n)}\right\}_{\nu=0}^\infty$, use the q-triangle inequality, and are done.

To prove inequality (4.1) for $\nu = 0$ is a simple matter. Note that for all $\nu \geq 0$

$$\left\langle \bar{f}_\nu, \phi \right\rangle = \int_R \langle f_\nu(\cdot, v), \phi \rangle \, \psi(v) \, d\mu(v) \quad \forall \phi \in \mathcal{S}(\mathbb{R}^n).$$

This follows easily from the definitions. Now each f_ν is an entire analytic function because of the Payley–Wiener–Schwartz theorem (cf. section 3.1). We estimate

$$(4.2) \ \|\bar{f}_0\|_{L_P(\mathbb{R}^n)} = \left\| \int_R f_0(\cdot, v) \psi(v) \, d\mu(v) \right\|_{L_P(\mathbb{R}^n)} \leq \mu(R)^{1/\rho_1'} \|\psi\|_F \|f_0\|_{L_P(\mathbb{R}^n, E_1)}$$

and then use the Nikol'skij inequality (3.4) with $\alpha = 0$ and $q = P$. Recall that R has finite μ-measure and that the exponent ρ_1 was defined by $1/\rho_1 = 1/r + 1/r_1$. This gives (4.1) for $\nu = 0$. So we will assume in the following that $\nu \geq 1$. Then both f_ν and g_ν are smooth functions, and the support of their Fourier transforms lies in a compact set strictly bounded away from the origin.

Since f and g satisfy the transport equation (2.1), the following identity holds:

$$(4.3) \qquad (i\mathbf{a}(v) \cdot \xi) \, \hat{f}(\xi, v) = \hat{g}(\xi, v) \quad \text{in } \mathcal{S}'(\mathbb{R}^n, E).$$

As explained in Theorem 3.1, the left-hand side (LHS) must be understood as a product of the symbol $i\mathbf{a} \cdot \xi \in \mathcal{O}_M(\mathbb{R}^n, L_\infty(R,\mu))$ with the tempered distribution $\hat{f} \in \mathcal{S}'(\mathbb{R}^n, E)$. More precisely, the LHS is that distribution for which

$$\left\langle (i\mathbf{a} \cdot \xi) \, \hat{f}, \phi \right\rangle = \sum_{j=1}^n \mathbf{a}_j \left\langle \hat{f}, i\xi_j \phi \right\rangle \quad \forall \phi \in \mathcal{S}(\mathbb{R}^n).$$

Note that multiplication with $\mathbf{a} \in L_\infty(R,\mu)$ maps E continuously into itself. Then identity (4.3) follows immediately from the definitions, and it still holds when we multiply both sides with φ_ν. So we may replace $\hat{f}$ and $\hat{g}$ with $\hat{f}_\nu$ and $\hat{g}_\nu$.

One might be tempted to divide by the symbol $i\mathbf{a} \cdot \xi$ and express $\hat{f}_\nu$ by $\hat{g}_\nu$. However, products of distributions are defined only for smooth functions. And since $(i\mathbf{a} \cdot \xi)^{-1}$ becomes unbounded for all $\xi \in \mathbb{R}^n$ orthogonal to $\mathbf{a}(v)$, we must be a little more careful. We introduce a splitting and use (4.3) only for that part of Fourier space that is bounded away from the set of points where the symbol vanishes. This is the classical approach. We choose an arbitrary nonnegative test function $\Pi \in \mathcal{D}(\mathbb{R})$, vanishing outside the interval $[-1,1]$, with $\Pi(\xi) = 1$ for all $|\xi| \le \frac{1}{2}$. Then we can define functions

$$\chi_{\mathbf{s}}(\xi,v) = \Pi\left(\delta^{-1}\mathbf{a}(v)\cdot\xi/|\xi|\right) \quad \text{and} \quad \chi_{\mathbf{r}}(\xi,v) = 1 - \chi_{\mathbf{s}}(\xi,v) \tag{4.4}$$

for $(\xi,v) \in \mathbb{R}^n \times R$ and $\delta \ge 0$. Note that the support of $\chi_{\mathbf{s}}$ is contained in the set

$$A_\delta = \{(\xi,v) \in \mathbb{R}^n \times R\colon |\mathbf{a}(v)\cdot\xi/|\xi|| \le \delta\}\,. \tag{4.5}$$

We claim now that both $\varphi_\nu\chi_{\mathbf{s}}$ and $\varphi_\nu\chi_{\mathbf{r}}(i\mathbf{a}\cdot\xi)^{-1}$ are bounded in $\mathcal{O}_M(\mathbb{R}^n, L_\infty(R,\mu))$ for any $\nu \ge 1$. If that is true, we obtain an identity

$$\hat{f}_\nu(\xi,v) = \chi_{\mathbf{s}}(\xi,v)\hat{f}_\nu(\xi,v) + \chi_{\mathbf{r}}(\xi,v)\frac{\hat{g}_\nu(\xi,v)}{i\mathbf{a}(v)\cdot\xi} \quad \text{in } \mathcal{S}'(\mathbb{R}^n,E) \tag{4.6}$$

for $\nu \ge 1$. In fact, note that the second term on the RHS of (4.6) is well defined in $\mathcal{S}'(\mathbb{R}^n,E)$ because of Theorem 3.1. Then we may use the relation (4.3) together with Remark 3.3 to eliminate $\hat{g}_\nu$, and (4.6) follows from the definition of $\chi_{\mathbf{s}}$ and $\chi_{\mathbf{r}}$.

To prove our claim, let us start with $\chi_{\mathbf{s}}$. This function is homogeneous of degree zero in ξ. Hence ξ-derivatives of it of order k are homogeneous in ξ of degree $-k$ for $k \ge 0$. More precisely, if $\alpha \in \mathbb{N}_0^n$ is some multi-index, $|\xi|^{|\alpha|}\partial_\xi^\alpha\chi_{\mathbf{s}}(\xi)$ is a linear combination of products of the following terms:

- derivatives of Π taken at $\delta^{-1}\mathbf{a}(v)\cdot\xi/|\xi|$,
- powers of $\delta^{-1}\mathbf{a}(v)\cdot\xi/|\xi|$ with positive exponent,
- polynomials in $\mathbf{a}(v) \in \mathbb{R}^n$, and
- polynomials in $\xi/|\xi| \in \mathbb{R}^n$.

This follows easily from an induction argument. We assumed that $\mathbf{a} \in L_\infty(R,\mu)$. Therefore, all of these terms are uniformly bounded with respect to v for any $\xi \in \mathbb{R}^n$ fixed. And since φ_ν vanishes in a neighborhood of zero for $\nu \ge 1$, the functions $\varphi_\nu\chi_{\mathbf{s}}$ are bounded in $\mathcal{O}_M(\mathbb{R}^n, L_\infty(R,\mu))$. We can proceed in the same manner for $\varphi_\nu\chi_{\mathbf{r}}$.

To prove that even $\varphi_\nu\chi_{\mathbf{r}}(i\mathbf{a}\cdot\xi)^{-1}$ is bounded in $\mathcal{O}_M(\mathbb{R}^n, L_\infty(R,\mu))$, note first that this function is homogeneous of degree -1 in ξ. Taking derivatives with respect to ξ, we obtain the same terms we already had for $\chi_{\mathbf{s}}$, but now there are also powers of $\mathbf{a}(v)\cdot\xi/|\xi|$ with negative exponent. Still, these terms are uniformly bounded in v because $\chi_{\mathbf{r}}$ vanishes in $A_{\delta/2}$ by construction. This proves our claim.

Summing up, we have a decomposition $\bar{f}_\nu = \bar{f}_{\mathbf{s},\nu} + \bar{f}_{\mathbf{r},\nu}$ for $\nu \ge 1$ with

$$\left\langle \bar{f}_{\mathbf{s},\nu}, \phi \right\rangle = \int_R \left\langle \left[\chi_{\mathbf{s}}\hat{f}_\nu\right](\cdot,v), \check{\phi} \right\rangle \psi(v)\,d\mu(v) \quad \text{and} \tag{4.7}$$

$$\left\langle \bar{f}_{\mathbf{r},\nu}, \phi \right\rangle = \int_R \left\langle \left[|\cdot|^{-1}\bar{\chi}_{\mathbf{r}}\hat{g}_\nu\right](\cdot,v), \check{\phi} \right\rangle \psi(v)\,d\mu(v) \quad \forall \phi \in \mathcal{S}(\mathbb{R}^n). \tag{4.8}$$

We put $\bar{\chi}_{\mathbf{r}}(\xi,v) = \chi_{\mathbf{r}}(\xi,v)\left(i\mathbf{a}(v)\cdot\xi/|\xi|\right)^{-1}$. While up to now the argument was the same for both cases of Theorem 2.2, we must now specialize a bit.

4.1. Regularity—Case I. We will use the fact that each $f \in B^s_{p,q}(\mathbb{R}^n, E)$ can be decomposed into an infinite sum of simple building blocks of the form $\lambda\mathbf{Q}$, where $\mathbf{Q}$ is a scalar function with nice properties and λ is an element of E.

4.1.1. The subatomic decomposition. Let $Q_{\nu m}$ be a cube in $\mathbb{R}^n$ centered at $2^{-\nu}m$ with sides parallel to the coordinate axes and side length $2^{-\nu}$, where $m \in \mathbb{Z}^n$ and $\nu \in \mathbb{N}_0$. If Q is a cube in $\mathbb{R}^n$ and $c > 1$ is a number, we will write cQ for the cube concentric with Q but with sides c times as long as those of Q. Now choose a nonnegative function $\psi \in \mathcal{S}(\mathbb{R}^n)$ with compact support in some cQ_{00} and

$$\sum_{m\in\mathbb{Z}^n} \psi(x-m) = 1 \quad \forall x \in \mathbb{R}^n. \tag{4.9}$$

Let $s \in \mathbb{R}, 0 < p \leq \infty, \frac{L+1}{2} \in \mathbb{N}_0, \gamma \in \mathbb{N}_0^n$, and $\psi^\gamma(x) = x^\gamma\psi(x)$. Then the function

$$(\gamma qu)^L_{\nu m}(x) = 2^{-\nu(s-\frac{n}{p})}\left((-\Delta)^{\frac{L+1}{2}}\psi^\gamma\right)(2^\nu x - m) \tag{4.10}$$

is called an $(s,p)_L - \gamma$-quark over the cube $Q_{\nu m}$.

Remark 4.1. This definition is taken from [22, section 14.1]. In the following, we will need only the case $L = -1$. Then the Laplace operator in (4.10) drops out.

Before stating the subatomic decomposition for $B^s_{p,q}(\mathbb{R}^n, E)$, we show the following.

THEOREM 4.2. *Let $K \subset \mathbb{R}^n$ be a compact set, and let $0 < p \leq \infty$. Then there exists a number $\kappa > 0$ with the following property: for any $\mu \in \mathbb{N}$ with $\mu > \kappa$ the entire analytic function $g \in L_{p,K}(\mathbb{R}^n, E)$ can be written as*

$$g(x) = \sum_{m\in\mathbb{Z}^n}\sum_{\gamma\in\mathbb{N}_0^n} \lambda^\gamma_m \psi^\gamma(2^\mu x - m), \tag{4.11}$$

where the sum converges in $L_q(\mathbb{R}^n, E)$ for all $q \in [p,\infty]$ with $\lambda^\gamma_m \in E$ s.t.

$$\sup_{\gamma\in\mathbb{N}_0^n} 2^{\mu|\gamma|}\left(\sum_{m\in\mathbb{Z}^n} \|\lambda^\gamma_m\|^p_E\right)^{1/p} \leq C\|g\|_{L_p(\mathbb{R}^n,E)}. \tag{4.12}$$

The constant $C = C(K,\mu,n,p)$ does not depend on g.

Proof. Our proof somewhat simplifies a similar argument in Triebel [22, section 14.15]. Choose $\rho \in \mathcal{S}(\mathbb{R}^n)$ with $\operatorname{supp}\hat{\rho}$ contained in some compact neighborhood of K and $\hat{\rho}(\xi) = 1$ for all $\xi \in K$. From the Paley–Wiener–Schwartz theorem we infer that ρ is an entire analytic function. Moreover,

$$g(x) = \int_{\mathbb{R}^n} g(y)\rho(x-y)\,dy \quad \forall x \in \mathbb{R}^n \tag{4.13}$$

(cf. section 3.2). Note that as a consequence of the Nikol'skij inequality (3.4) the function g is bounded and hence locally integrable. We now expand $\rho(\cdot - y)$ into a Taylor series around the point $2^{-\mu}m$ with $m \in \mathbb{Z}^n$ and $\mu \in \mathbb{N}$ fixed. Then

$$\begin{aligned}\psi(2^\mu x - m)\rho(x-y) &= \psi(2^\mu x - m)\sum_{\gamma\in\mathbb{N}_0^n}\frac{D^\gamma\rho(2^{-\mu}m - y)}{\gamma!}\left(x - 2^{-\mu}m\right)^\gamma \\ &= \sum_{\gamma\in\mathbb{N}_0^n} 2^{-\mu|\gamma|}\frac{D^\gamma\rho(2^{-\mu}m-y)}{\gamma!}\psi^\gamma(2^\mu x - m). \end{aligned} \tag{4.14}$$

Since $\operatorname{supp}\psi \subset cQ_{00}$ with $c > 1$, we have $|\psi^\gamma(2^\mu x - m)| \leq (\frac{c}{2}\sqrt{2}^n)^{|\gamma|}$. We apply the Cauchy integral formula componentwise to ρ. Then we obtain for all $z \in \mathbb{C}^n$

$$\rho(z_1, \dots, z_n) = (2\pi i)^{-n} \int_{|w_1 - z_1| = 1} \cdots \int_{|w_n - z_n| = 1} \frac{\rho(w_1, \dots, w_n)\, dw_1 \cdots dw_n}{(z_1 - w_1) \cdots (z_n - w_n)}$$

(integration over $\mathbb{C}^n$). Differentiation gives

$$\begin{aligned} &D^\gamma \rho(z_1, \dots, z_n) \\ &= (-1)^{|\gamma|} \gamma! (2\pi i)^{-n} \int_{|w_1 - z_1| = 1} \cdots \int_{|w_n - z_n| = 1} \frac{\rho(w_1, \dots, w_n)\, dw_1 \cdots dw_n}{(z_1 - w_1)^{\gamma_1 + 1} \cdots (z_n - w_n)^{\gamma_n + 1}} \end{aligned}$$

for $z \in \mathbb{C}^n$ and $\gamma \in \mathbb{N}_0^n$. Using (3.3), we now obtain for arbitrary $N \in \mathbb{N}$ the estimate

$$|D^\gamma \rho(z)| \leq C_N \gamma! (2\pi)^{-n} \int_{|w_1 - z_1| = 1} \cdots \int_{|w_n - z_n| = 1} (1 + |w|)^{-N} e^{c|\mathrm{Im} w|}\, dw_1 \cdots dw_n.$$

If $z \in \mathbb{R}^n$, then $|\mathrm{Im} w| \leq 1$ in the domain of integration. Using

$$1 + |z| \leq (1 + |w|)(1 + |z - w|) \leq (1 + |w|)(1 + |z_1 - w_1| + \cdots + |z_n - w_n|),$$

we can find some constant $C = C(K, N, n)$ s.t.

$$|D^\gamma \rho(z)| \leq C \gamma! (1 + |z|)^{-N} \quad \forall z \in \mathbb{R}^n,\ \gamma \in \mathbb{N}_0^n. \tag{4.15}$$

The number of multi-indices $\gamma \in \mathbb{N}_0^n$ with $|\gamma| = k$ grows polynomially in k. We conclude that the expansion (4.14) is absolutely convergent for μ large enough:

$$\begin{aligned} &\left| \psi(2^\mu x - m) \rho(x - y) - \sum_{|\gamma| \leq K} 2^{-\mu|\gamma|} \frac{D^\gamma \rho(2^{-\mu} m - y)}{\gamma!} \psi^\gamma(2^\mu x - m) \right| \\ &\leq C \mathbf{1}_{cQ_{\mu m}}(x) \left(1 + |2^{-\mu} m - y|\right)^{-N} \underbrace{\sum_{k = K+1}^\infty k^n 2^{-\mu k} \left(\frac{c}{2}\sqrt{2}^n\right)^k}_{c_K} \end{aligned} \tag{4.16}$$

and $c_K \to 0$ for $K \to \infty$. This estimate implies (4.11) with

$$\lambda_m^\gamma = 2^{-\mu|\gamma|} \int_{\mathbb{R}^n} g(y) \frac{(D^\gamma \rho)(2^{-\mu} m - y)}{\gamma!} dy \in E. \tag{4.17}$$

To see that, note first that the sum in m does not cause any harm: For $x \in \mathbb{R}^n$ fixed only finitely, many terms contribute to the sum because ψ is compactly supported. Using (4.9), we can write for any $x \in \mathbb{R}^n$

$$\begin{aligned} &\left\| g(x) - \sum_{m \in \mathbb{Z}^n} \sum_{|\gamma| \leq K} \lambda_m^\gamma \psi^\gamma(2^\mu x - m) \right\|_E \\ &\leq \sum_{m \in \mathbb{Z}^n} \left\| \psi(2^\mu x - m) g(x) - \sum_{|\gamma| \leq K} \lambda_m^\gamma \psi^\gamma(2^\mu x - m) \right\|_E . \end{aligned}$$

Consider one single term of this sum. If $g(x)$ is replaced with (4.13) and λ_m^γ with (4.17), then we obtain the following:

$$\left\|\int_{\mathbb{R}^n} g(y)\left(\psi(2^\mu x-m)\rho(x-y)\right.\right.$$
$$\left.\left.-\sum_{|\gamma|\le K}2^{-\mu|\gamma|}\frac{(D^\gamma\rho)\,(2^{-\mu}m-y)}{\gamma!}\psi^\gamma(2^\mu x-m)\right)dy\right\|_E.$$

The integrand is smooth and hence strongly measurable: g is an entire analytic function, and the terms in brackets are in $\mathcal{S}(\mathbb{R}^n)$. Note that the sum in γ is finite here. Then the Bochner theorem allows us to push the E-norm inside the integral. We apply (4.16) and estimate

$$\int_{R^n}\|g(y)\|_E\left|\psi(2^\mu x-m)\rho(x-y)\right.$$
$$\left.-\sum_{|\gamma|\le K}2^{-\mu|\gamma|}\frac{(D^\gamma\rho)\,(2^{-\mu}m-y)}{\gamma!}\psi^\gamma(2^\mu x-m)\right|dy$$
$$\le Cc_K\mathbf{1}_{cQ_{\mu m}}(x)\int_{\mathbb{R}^n}\|g(y)\|_E\left(1+|2^{-\mu}m-y|\right)^{-N}dy.$$

Recall that N can be made arbitrarily large. Moreover,

$$1+|x-y|\le\left(1+|x-2^{-\mu}m|\right)\left(1+|2^{-\mu}m-y|\right)\le\left(1+c2^{-\mu}\right)\left(1+|2^{-\mu}m-y|\right)$$

for all $x\in cQ_{\mu m}$. Using Theorem 3.5, we find some constant $C>0$ s.t. for $w<p$

$$\int_{\mathbb{R}^n}\|g(y)\|_E\left(1+|2^{-\mu}m-y|\right)^{-N}dy$$
$$\le C\sup_{y\in\mathbb{R}^n}\|g(y)\|_E\,(1+|x-y|)^{-\frac{n}{w}}\cdot\int_{\mathbb{R}^n}(1+|x-y|)^{-N+\frac{n}{w}}\,dy$$
$$\le C\left(\mathbf{M}\|g\|_E^w\right)^{1/w}(x)\quad\forall x\in cQ_{\mu m}. \tag{4.18}$$

The constant $C=C(\mu,n,p,w,K,N)$ does not depend on g. We obtain

$$\left\|g(x)-\sum_{m\in\mathbb{Z}^n}\sum_{|\gamma|\le K}\lambda_m^\gamma\psi^\gamma(2^\mu x-m)\right\|_E\le Cc_K\left(\mathbf{M}\|g\|_E^w\right)^{1/w}(x)\cdot\sum_{m\in\mathbb{Z}^n}\mathbf{1}_{cQ_{\mu m}}(x)$$

for all $x\in\mathbb{R}^n$. Note that the m-sum on the RHS is uniformly bounded because only finitely many cubes $cQ_{\mu m}$ overlap. Now we take $L_q(\mathbb{R}^n)$ (quasi) norms on both sides. Since $w<p\le q$ (hence $q/w>1$) we apply the Hardy–Littlewood maximal inequality (cf. Stein [20]) and obtain

$$\left\|g(x)-\sum_{m\in\mathbb{Z}^n}\sum_{|\gamma|\le K}\lambda_m^\gamma\psi^\gamma(2^\mu x-m)\right\|_{L_q(\mathbb{R}^n,E)}\le Cc_K\left\|\mathbf{M}\|g\|_E^w\right\|_{L_{q/w}(\mathbb{R}^n)}^{1/w}$$
$$\text{and}\quad\left\|\mathbf{M}\|g\|_E^w\right\|_{L_{q/w}(\mathbb{R}^n)}^{1/w}\le C\|g\|_{L_q(\mathbb{R}^n,E)}\le C\|g\|_{L_p(\mathbb{R}^n,E)}.$$

In the last step we used the Nikol'skij inequality (3.4). Now c_K vanishes if $K \to \infty$. Therefore, (4.11) converges strongly in $L_q(\mathbb{R}^n, E)$ for any $q \geq p$ as claimed.

To prove (4.12) we need only to modify this argument a little. Note first that the $\ell_p(\mathbb{Z}^n)$ (quasi) norm in m can also be realized like this:

$$\begin{aligned}\sum_{m\in\mathbb{Z}^n} \|\lambda_m^\gamma\|_E^p &= \sum_{m\in\mathbb{Z}^n} \|\lambda_m^\gamma\|_E^p \cdot 2^{\mu n} \int_{\mathbb{R}^n} \mathbf{1}_{Q_{\mu m}}(x)\,dx \\ &= 2^{\mu n} \int_{\mathbb{R}^n} \sum_{m\in\mathbb{Z}^n} \|\lambda_m^\gamma\|_E^p \mathbf{1}_{Q_{\mu m}}(x)\,dx \\ &= 2^{\mu n} \int_{\mathbb{R}^n} \left(\sum_{m\in\mathbb{Z}^n} \|\lambda_m^\gamma\|_E \mathbf{1}_{Q_{\mu m}}(x) \right)^p dx\end{aligned}$$

because the $Q_{\mu m}$ are pairwise disjoint. On the other hand, we find with (4.15)

$$2^{\mu|\gamma|} \|\lambda_m^\gamma\|_E \leq C \int_{\mathbb{R}^n} \|g(y)\|_E \left(1 + |2^{-\mu}m - y|\right)^{-N} dy.$$

If we now continue with (4.18), we obtain (4.12). The proof is complete. □

Now we can present the subatomic decomposition for $B^s_{p,q}(\mathbb{R}^n, E)$.

THEOREM 4.3. *Let $0 < p, q \leq \infty$, and $s > \sigma_p = \max\{n\,(1/p - 1)\,, 0\}$. Then there exists a number $\kappa > 0$ s.t. for any $\mu \in \mathbb{N}$ with $\mu > \kappa$ any $f \in B^s_{p,q}(\mathbb{R}^n, E)$ can be decomposed into an infinite sum*

$$f(x) = \sum_{\nu=0}^{\infty} \sum_{m\in\mathbb{Z}^n} \sum_{\gamma\in\mathbb{N}_0^n} \lambda^\gamma_{\nu m} \mathbf{Q}^\gamma_{\nu m}(x) \qquad \text{in } \mathcal{S}'(\mathbb{R}^n, E). \tag{4.19}$$

The $\mathbf{Q}^\gamma_{\nu m}$ are $(s,p)_{-1} - \gamma$-quarks and the $\lambda^\gamma_{\nu m} \in E$ coefficients with

$$\sup_{\gamma\in\mathbb{N}_0^n} 2^{\mu|\gamma|} \left(\sum_{\nu=0}^{\infty} \left(\sum_{m\in\mathbb{Z}^n} \|\lambda^\gamma_{\nu m}\|_E^p \right)^{q/p} \right)^{1/q} < \infty. \tag{4.20}$$

Vice versa, if coefficients $\lambda^\gamma_{\nu m} \in E$ are given with (4.20)*, then the sum* (4.19) *converges in $\mathcal{S}'(\mathbb{R}^n, E)$ and defines an element in $B^s_{p,q}(\mathbb{R}^n, E)$. The* inf *in* (4.20) *over all admissible representations* (4.19) *is an equivalent (quasi) norm in $B^s_{p,q}(\mathbb{R}^n, E)$.*

Remark 4.4. This is Corollary 15.9 in Triebel [22].

One direction of the proof is straightforward: For $f \in \mathcal{S}'(\mathbb{R}^n, E)$ we apply Theorem 4.2 to the dyadic parts f_ν. Let us consider the family $g_\nu(y) = f_\nu(2^{-\nu}y)$ for $y \in \mathbb{R}^n$. From the definition of the Fourier transform we easily find that $\operatorname{supp} \hat{g}_\nu \subset B_2(0)$ for all $\nu \in \mathbb{N}_0$. Hence the supports of $\hat{g}_\nu$ are all contained in one single fixed compact subset $K \subset \mathbb{R}^n$. Then there exists some $\kappa > 0$ s.t. for each $\mu > \kappa$ and all $\nu \in \mathbb{N}_0$ the entire analytic function g_ν can be decomposed into

$$g_\nu(y) = \sum_{m\in\mathbb{Z}^n} \sum_{\gamma\in\mathbb{N}_0^n} \tilde{\lambda}^\gamma_{\nu m} \psi^\gamma(2^\mu y - m),$$

with strong convergence in $L^q(\mathbb{R}^n, E)$ for all $q \geq p$, and coefficients $\tilde{\lambda}^\gamma_{\nu m} \in E$ with

$$\sup_{\gamma\in\mathbb{N}_0^n} 2^{\mu|\gamma|} \left(\sum_{m\in\mathbb{Z}^n} \|\tilde{\lambda}^\gamma_{\nu m}\|_E^p \right)^{1/p} \leq C\|g_\nu\|_{L_p(\mathbb{R}^n,E)} = C 2^{\nu\frac{n}{p}} \|f_\nu\|_{L_p(\mathbb{R}^n,E)}. \tag{4.21}$$

The constant C can be chosen independent of ν. Renormalizing, we can write

$$f_\nu(x) = g_\nu(2^\nu x) = 2^{\mu(s-\frac{n}{p})} \sum_{m\in\mathbb{Z}^n} \sum_{\gamma\in\mathbb{N}_0^n} \underbrace{2^{\nu(s-\frac{n}{p})}\tilde{\lambda}^\gamma_{\nu m}}_{=\lambda^\gamma_{\nu m}} \underbrace{(\gamma q u)^{-1}_{\mu+\nu,m}(x)}_{=\mathbf{Q}^\gamma_{\nu m}(x)} \tag{4.22}$$

for $x \in \mathbb{R}^n$. This is (4.19). Now the proof of (4.20) is easy. From (4.21) we obtain

$$\begin{aligned}
&\sup_{\gamma\in\mathbb{N}_0^n} 2^{\mu|\gamma|} \left(\sum_{\nu=0}^\infty \left(\sum_{m\in\mathbb{Z}^n} \|\lambda^\gamma_{\nu m}\|^p_E \right)^{q/p} \right)^{1/q} \\
&\quad \le C \left(\sum_{\nu=0}^\infty \left(2^{\nu s} 2^{-\nu\frac{n}{p}} \sup_{\gamma\in\mathbb{N}_0^n} 2^{\mu|\gamma|} \left(\sum_{m\in\mathbb{Z}^n} \|\tilde{\lambda}^\gamma_{\nu m}\|^p_E \right)^{1/p} \right)^q \right)^{1/q} \\
&\quad \le C \left(\sum_{\nu=0}^\infty 2^{\nu s q} \|f_\nu\|^q_{L_p(\mathbb{R}^n,E)} \right)^{1/q} = C\|f\|_{B^s_{p,q}(\mathbb{R}^n,E)}.
\end{aligned}$$

We used the Minkowski inequality. Proving the reverse direction is more elaborate, and we do not want to give the details here. We refer again to Triebel [22].

4.1.2. Proof of Theorem 2.2, Case I. Let us consider first the term $\bar{f}_{\mathbf{s},\nu}$ in (4.7). If we assume for the moment that $f_\nu(x,v) = \lambda(v)\mathbf{Q}(x)$ for suitable $\mathbf{Q} \in L_2(\mathbb{R}^n)$ and $\lambda \in E_1$, then $\hat{f}_\nu = \lambda\hat{\mathbf{Q}}$ is a measurable function, and we find

$$\langle \bar{f}_{\mathbf{s},\nu}, \phi \rangle = \int_{\mathbb{R}^n} M(\xi)\hat{\mathbf{Q}}(\xi)\check{\phi}(\xi)\,d\xi \quad \text{with} \quad M(\xi) = \int_R \chi_{\mathbf{s}}(\xi,v)\,(\psi\lambda)\,(v)\,d\mu(v).$$

In that situation, we obtain $\bar{f}_{\mathbf{s},\nu}$ by simply applying the Fourier multiplier operator M to $\mathbf{Q}$. We can use Plancherel's theorem to estimate

$$\left|\langle \bar{f}_{\mathbf{s},\nu}, \phi \rangle\right| \le \|M\|_{L_\infty(\mathbb{R}^n)} \|\mathbf{Q}\|_{L_2(\mathbb{R}^n)} \|\phi\|_{L_2(\mathbb{R}^n)}.$$

By assumption, the product $\psi\lambda$ is bounded in $L_{\rho_1}(R,\mu)$. Therefore, we may use the Hölder inequality to estimate

$$\|M\|_{L_\infty(\mathbb{R}^n)} = \sup_{\xi\in\mathbb{R}^n} \left| \int_R \chi_{\mathbf{s}}(\xi,v)\,(\psi\lambda)\,(v)\,d\mu(v) \right| \le \|\psi\|_F \|\lambda\|_{E_1} \sup_{\xi\in\mathbb{R}^n} \|\mathbf{1}_{A_\delta(\xi)}\|_{L_{\rho_1'}(R,\mu)},$$

where $A_\delta(\xi) = \{v \in R\colon (\xi,v) \in A_\delta\}$. However, the nondegeneracy condition (2.4) for the velocity field $\mathbf{a}$ bounds the measure of the set A_δ uniformly in $\xi \in \mathbb{R}^n$. Therefore,

$$\|\bar{f}_{\mathbf{s},\nu}\|_{L_2(\mathbb{R}^n)} = \sup_{\phi\in\mathcal{S}(\mathbb{R}^n)} \|\phi\|^{-1}_{L_2(\mathbb{R}^n)} \left|\langle \bar{f}_{\mathbf{s},\nu}, \phi \rangle\right| \le C\delta^{\alpha/\rho_1'} \|\psi\|_F \|\lambda\|_{E_1} \|\mathbf{Q}\|_{L_2(\mathbb{R}^n)}. \tag{4.23}$$

The conclusion is that $\bar{f}_{\mathbf{s},\nu}$ becomes small in $L_2(\mathbb{R}^n)$ for $\delta \to 0$ if $\mathbf{Q} \in L_2(\mathbb{R}^n)$, $\lambda \in E_1$, and $\rho_1 > 1$. If $\rho_1 = 1$, the estimate does not depend on δ anymore.

We now consider the second term $\bar{f}_{\mathbf{r},\nu}$ in (4.8). If we assume again for the moment that $g_\nu(x,v) = \lambda(v)\mathbf{Q}(x)$ for suitable $\mathbf{Q} \in L_2(\mathbb{R}^n)$ and $\lambda \in E_2$, then $\hat{g}_\nu = \lambda\hat{\mathbf{Q}}$ is a measurable function, and we can write

$$\langle \bar{f}_{\mathbf{r},\nu}, \phi \rangle = \int_{\operatorname{supp}\varphi_\nu} |\xi|^{-1} \bar{M}(\xi)\hat{\mathbf{Q}}(\xi)\check{\phi}(\xi)\,d\xi \quad \text{with } \bar{M}(\xi) = \int_R \bar{\chi}_{\mathbf{r}}(\xi,v)\,(\psi\lambda)\,(v)\,d\mu(v).$$

To estimate $\bar{f}_{\mathbf{r},\nu}$ in $L_2(\mathbb{R}^n)$, we need to find an L_∞-bound for the multiplier. We have $|\xi|^{-1} \le c2^{-\nu}$ in the domain of integration by the construction of φ_ν. Moreover,

$$\|\bar{M}\|_{L_\infty(\mathbb{R}^n)} \le \|\psi\|_F \|\lambda\|_{E_2} \sup_{\xi\in\mathbb{R}^n} \left\| (i\mathbf{a}\cdot\xi/|\xi|)^{-1} \mathbf{1}_{R\setminus A_{\delta/2}(\xi)} \right\|_{L_{\rho_2'}(R,\mu)}.$$

Now we can use the following result.

Lemma 4.5. *Assume* (2.4). *Then for every $\rho \ge 1$ with $\alpha < \rho$*

$$\sup_{\xi\in\mathbb{R}^n} \int_{R\setminus A_\delta(\xi)} |i\mathbf{a}(v)\cdot\xi/|\xi||^{-\rho}\, d\mu(v) \le C\delta^{\alpha-\rho}.$$

Remark 4.6. Here $C = C(\alpha,\rho)$ does not depend on δ. Estimates of this kind appear in many papers on velocity averaging, e.g., in [13], [16], [5].

We use Lemma 4.5 with $\rho = \rho_2'$ under the assumption that $\alpha/\rho_2' < 1$. By testing $\bar{f}_{\mathbf{r},\nu}$ against all functions $\phi \in \mathcal{S}(\mathbb{R}^n)$, we obtain

$$\|\bar{f}_{\mathbf{r},\nu}\|_{L_2(\mathbb{R}^n)} \le C2^{-\nu}\delta^{-1+\alpha/\rho_2'}\|\psi\|_F\|\lambda\|_{E_2}\|\mathbf{Q}\|_{L_2(\mathbb{R}^n)}.$$

We conclude that the dyadic elements $\bar{f}_{\mathbf{r},\nu}$ of the average vanish in $L_2(\mathbb{R}^n)$ like $2^{-\nu}$ if $\nu\to\infty$. This corresponds to a gain of regularity of one derivative (cf. Definition 3.7). Note, however, that $\delta^{-1+\alpha/\rho_2'}$ becomes large as $\delta\to 0$.

Now we use Theorem 4.2, which tells us that the dyadic blocks of f_ν and g_ν can be realized as tensor products. Recall (4.22) from section 4.1.1: For suitable coefficients $\lambda_{\nu m}^\gamma \in E_1$ and $(s_1,p)_{-1} - \gamma$-quarks $\mathbf{Q}_{\nu m}^\gamma$, we have

$$f_\nu(x,v) = 2^{\mu(s-\frac{n}{p})} \sum_{m\in\mathbb{Z}^n}\sum_{\gamma\in\mathbb{N}_0^n} \lambda_{\nu m}^\gamma(v)\mathbf{Q}_{\nu m}^\gamma(x).$$

To control $\bar{f}_{\mathbf{s},\nu}$ in $L_2(\mathbb{R}^n)$, we can now use the triangle inequality and obtain

$$\|\bar{f}_{\mathbf{s},\nu}\|_{L_2(\mathbb{R}^n)} \le C\delta^{\alpha/\rho_1'}\|\psi\|_F \sum_{m\in\mathbb{Z}^n}\sum_{\gamma\in\mathbb{N}_0^n} \|\lambda_{\nu m}^\gamma\|_{E_1}\|\mathbf{Q}_{\nu m}^\gamma\|_{L_2(\mathbb{R}^n)}. \tag{4.24}$$

Here the constant $C = C(\mu,n,p)$ does not depend on f or ν. Note now that quarks are normalized, i.e., there exists a constant $C = C(\mu)$ s.t.

$$\|\mathbf{Q}_{\nu m}^\gamma\|_{L_2(\mathbb{R}^n)} \le C2^{\sigma|\gamma|}2^{-\nu s_1+\nu n(\frac{1}{p}-\frac{1}{2})} \qquad \text{for some } \sigma>0 \text{ and } \gamma\in\mathbb{N}_0^n, m\in\mathbb{Z}^n.$$

One nice feature of the subatomic decomposition in Theorem 4.2 is that we can choose the parameter $\mu\in\mathbb{N}$ as large as we want (at the expense of enlarging the constants, of course). For suitable μ and with $\kappa = n(\frac{1}{p}-\frac{1}{2})$, we therefore obtain the following estimate:

$$\begin{aligned}
\|\bar{f}_{\mathbf{s},\nu}\|_{L_2(\mathbb{R}^n)} &\le C\delta^{\alpha/\rho_1'}2^{-\nu s_1}2^{\nu\kappa}\|\psi\|_F \\
&\qquad\cdot \sup_{\gamma\in\mathbb{N}_0^n} 2^{\mu|\gamma|}\left(\sum_{m\in\mathbb{Z}^n}\|\lambda_{\nu m}^\gamma\|_{E_1}^p\right)^{1/p} \cdot \sum_{\gamma\in\mathbb{N}_0^n} 2^{-(\mu-\sigma)|\gamma|} \\
&\le C\delta^{\alpha/\rho_1'}2^{\nu\kappa}\|\psi\|_F\|f_\nu\|_{L_p(\mathbb{R}^n,E_1)}. \qquad (4.25)
\end{aligned}$$

Consult (4.21)–(4.22). Here we used the fact that $\ell_p(\mathbb{Z}^n) \hookrightarrow \ell_1(\mathbb{Z}^n)$ if $p\le 1$. A similar argument works for $\bar{f}_{\mathbf{r},\nu}$ with $\nu\ge 1$. We have

$$\|\bar{f}_{\mathbf{r},\nu}\|_{L_2(\mathbb{R}^n)} \le C\delta^{-1+\alpha/\rho_2'}2^{-\nu}2^{\nu\kappa}\|\psi\|_F\|g_\nu\|_{L_p(\mathbb{R}^n,E_2)}.$$

The rest of the proof is only a matter of matching: For all $\nu \geq 1$, we want

$$\begin{aligned} &2^{\nu S}\|\bar{f}_{\mathbf{s},\nu}\|_{L_2(\mathbb{R}^n)} \leq C\|\psi\|_F \cdot 2^{\nu s}\|f_\nu\|_{L_p(\mathbb{R}^n,E_1)} \quad \text{and} \\ &2^{\nu S}\|\bar{f}_{\mathbf{r},\nu}\|_{L_2(\mathbb{R}^n)} \leq C\|\psi\|_F \cdot 2^{\nu(s-\tau)}\|g_\nu\|_{L_p(\mathbb{R}^n,E_2)}. \end{aligned} \tag{4.26}$$

To provide this, we choose the ansatz $\delta = 2^{-\nu\sigma}$ and solve the system of equations

$$\begin{aligned} S - \sigma\frac{\alpha}{\rho_1'} &= s - \kappa, \\ S - \sigma\left(-1 + \frac{\alpha}{\rho_2'}\right) &= 1 + s - \tau - \kappa \end{aligned}$$

for (σ, S). We find

$$\begin{aligned} \sigma &= (1-\tau)\left[1 + \alpha\left(\frac{1}{\rho_2} - \frac{1}{\rho_1}\right)\right]^{-1} \quad \text{and} \\ S &= s - n\left(\frac{1}{p} - \frac{1}{2}\right) + (1-\tau)\frac{\alpha}{\rho_1'}\left[1 + \alpha\left(\frac{1}{\rho_2} - \frac{1}{\rho_1}\right)\right]^{-1}. \end{aligned}$$

S is just the regularity of the average $\bar{f}$ we seek. Note also that $\sigma \geq 0$ and hence $\delta \leq 1$ because $\tau \leq 1$. Using (4.26), we obtain the estimate (4.1) for all $\nu \geq 1$.

4.2. Regularity—Case II.

To prove estimate (4.1) for $1 < p < \frac{n}{n-1}$, we use the fact that the action of some homogeneous Fourier multiplier operator on smooth functions can be rewritten in terms of the well-known Radon transform.

4.2.1. Some remarks on the Radon transform.

The Radon transform $\mathbf{R}$ maps a function $\Phi \in \mathcal{S}(\mathbb{R}^n)$ to the average of Φ over all $n-1$-dimensional hyperplanes in $\mathbb{R}^n$. Every such hyperplane is characterized by (1) the unit normal vector $\omega \in S^{n-1}$ and (2) the distance $r \geq 0$ between hyperplane and origin. We therefore define

$$\mathbf{R}\Phi(\omega, r) = \int_{\omega\cdot x = r} \Phi(x)\,dS(x) \qquad \text{for } (\omega, r) \in S^{n-1} \times [0,\infty). \tag{4.27}$$

Here dS is the induced Lebesgue measure. As a synonym we will also write $\tilde{\Phi} = \mathbf{R}\Phi$. Putting $\tilde{\Phi}(\omega, r) = \tilde{\Phi}(-\omega, -r)$, the Radon transform can be extended to a function on $S^{n-1} \times \mathbb{R}$. We have the following relationship with the Fourier transform:

$$\begin{aligned} (2\pi)^n\check{\Phi}(s\omega) &= \int_{\mathbb{R}^n} e^{is\omega\cdot x}\Phi(x)\,dx = \int_{\mathbb{R}} e^{isr}\left(\int_{\omega\cdot x = r}\Phi(x)\,dS(x)\right)dr \\ &= \int_{\mathbb{R}} e^{isr}\tilde{\Phi}(\omega, r)\,dr = 2\pi\mathbf{F}^{-1}\tilde{\Phi}(\omega, s) \qquad \text{for } (\omega, s) \in S^{n-1} \times \mathbb{R}. \end{aligned} \tag{4.28}$$

Let $m \in L_\infty(\mathbb{R}^n)$ be an even homogeneous function of degree zero. For simplicity we assume $m \in C^\infty(S^{n-1})$. Choose $\varphi \in \mathcal{S}(\mathbb{R}^n)$ with $\operatorname{supp}\check{\varphi} \subset \mathbb{R}^n \backslash B_1(0)$. Then

$$\begin{aligned} \mathbf{F}[m\check{\varphi}](x) &= \int_{\mathbb{R}^n} e^{-ix\cdot\xi} m(\xi)\check{\varphi}(\xi)\,d\xi \\ &= \int_{S^{n-1}} m(\omega)\int_0^\infty e^{-ix\cdot s\omega}s^{n-1}\check{\varphi}(s\omega)\,ds\,d\omega. \end{aligned} \tag{4.29}$$

Note that $m\check{\varphi} \in \mathcal{S}(\mathbb{R}^n)$. Now we define $\Phi \in \mathcal{S}(\mathbb{R}^n)$ by $\check{\Phi}(\xi) = |\xi|^{n-1}\check{\varphi}(\xi)$ for all $\xi \in \mathbb{R}^n$. Since m is even, we obtain after a substitution $t = -s$ and $\sigma = -\omega$

$$\int_{S^{n-1}} m(\omega) \int_0^\infty e^{-ix\cdot s\omega} \check{\Phi}(s\omega)\, ds\, d\omega = \int_{S^{n-1}} m(\sigma) \int_{-\infty}^0 e^{-ix\cdot t\sigma} \check{\Phi}(t\sigma)\, dt\, d\sigma.$$

We may therefore extend the s-integral in (4.29) to the whole real line if we allow for an extra factor 1/2. If we now use equality (4.28), we find

$$\begin{aligned} \mathbf{F}[m\check{\varphi}](x) &= \frac{1}{2(2\pi)^{n-1}} \int_{S^{n-1}} m(\omega) \int_{\mathbb{R}} e^{-isx\cdot\omega} \mathbf{F}^{-1}\tilde{\Phi}(\omega, s)\, ds\, d\omega \\ &= \frac{1}{2(2\pi)^{n-1}} \int_{S^{n-1}} m(\omega) \tilde{\Phi}(\omega, x\cdot\omega)\, d\omega. \end{aligned} \tag{4.30}$$

The RHS is simply the average of $\tilde{\Phi}$ over all hyperplanes containing a given point $x \in \mathbb{R}^n$. Now $\mathbf{F}[|\cdot|^{n-1}\check{\varphi}]$ is just a power of the Laplacian, $\Delta^{\frac{n-1}{2}}\varphi$, times some constant (cf. Stein [20]). Hence we obtain for $m(\xi) = 1$ the following identity:

$$\varphi(x) = c\Delta^{\frac{n-1}{2}} \int_{S^{n-1}} \tilde{\varphi}(\omega, x\cdot\omega)\, d\omega.$$

This gives an inversion formula for the Radon transform.

4.2.2. Proof of Theorem 2.2, Case II. Consider first $\hat{f}_{\mathbf{s},\nu}$ in (4.7). Following the proof of Theorem 3.1, we obtain for $\chi_{\mathbf{s}}\hat{f}_\nu \in \mathcal{S}'(\mathbb{R}^n, E)$ and $\phi \in \mathcal{S}(\mathbb{R}^n)$

$$\left\langle \chi_{\mathbf{s}}\hat{f}_\nu, \check{\phi} \right\rangle = \lim_{\epsilon\to 0} \left\langle \chi_{\mathbf{s}}\hat{f}_{\nu,\epsilon}, \check{\phi} \right\rangle \quad \text{with } \hat{f}_{\nu,\epsilon} = \hat{f}_\nu \star \eta_\epsilon,$$

with η_ϵ as in section 3.1. Since $\hat{f}_\nu = \varphi_\nu \hat{f}$ is compactly supported, $\hat{f}_{\nu,\epsilon} \in \mathcal{S}(\mathbb{R}^n, E)$. We choose now some $\rho_1 \in \mathcal{S}(\mathbb{R}^n)$ with $\operatorname{supp}\rho_1 \subset B_5(0)\backslash B_{1/5}(0)$ and $\rho_1(\xi) = 1$ for $\xi \in \operatorname{supp}\varphi_1$, and we put $\rho_\nu(\xi) = \rho_1(2^{-\nu+1}\xi)$ for $\nu \geq 1$. Then we have for all $\epsilon > 0$

$$\left\langle \chi_{\mathbf{s}}\hat{f}_{\nu,\epsilon}, \check{\phi} \right\rangle = \int_{\mathbb{R}^n} \left(\chi_{\mathbf{s}}\rho_\nu\check{\phi}\right)(\xi) \cdot \hat{f}_{\nu,\epsilon}(\xi)\, d\xi = \int_{\mathbb{R}^n} \mathbf{F}[\chi_{\mathbf{s}}\rho_\nu\check{\phi}](x) \cdot \check{\eta}(\epsilon x) f_\nu(x)\, dx$$

in E. Note here that $f_\nu \in \mathcal{O}_M(\mathbb{R}^n, E)$ and $\mathbf{F}[\chi_{\mathbf{s}}\rho_\nu\check{\phi}] \in \mathcal{S}(\mathbb{R}^n, L_\infty(R,\mu))$. Using the dominated convergence theorem, we may let $\epsilon \to 0$ to find

$$\left\langle \chi_{\mathbf{s}}\hat{f}_\nu, \check{\phi} \right\rangle = \int_{\mathbb{R}^n} \mathbf{F}[\chi_{\mathbf{s}}\rho_\nu\check{\phi}](x) \cdot f_\nu(x)\, dx \quad \forall\phi \in \mathcal{S}(\mathbb{R}^n).$$

Now let us consider averages again. Since $\chi_{\mathbf{s}}(\xi, v)$ for fixed $v \in R$ is even and homogeneous in ξ, we can use (4.30). We obtain with $\Phi_\nu = \mathbf{F}[|\cdot|^{n-1}\rho_\nu\check{\phi}]$, $\nu \geq 1$,

$$\langle \bar{f}_{\mathbf{s},\nu}, \phi\rangle = \frac{1}{2(2\pi)^{n-1}} \int_R \int_{\mathbb{R}^n} \left(\int_{S^{n-1}} \chi_{\mathbf{s}}(\omega, v)\tilde{\Phi}_\nu(\omega, x\cdot\omega)\, d\omega \right) f_\nu(x, v)\psi(v)\, dx\, d\mu(v). \tag{4.31}$$

We must show that the integral exists (absolute integrability). We first do the integration in v. Then the Hölder inequality and our assumptions give us an estimate

$$\int_R |\psi(v) f_\nu(x, v)\chi_{\mathbf{s}}(\omega, v)|\, d\mu(v) \leq \|\psi\|_F \|f_\nu(x)\|_{E_1} \|\chi_{\mathbf{s}}(\omega)\|_{L_{\rho_1'}(R,\mu)} \tag{4.32}$$
$$\text{for } (x, \omega) \in \mathbb{R}^n \times S^{n-1}.$$

Since $\chi_{\mathbf{s}}(\omega, v) \leq \mathbf{1}_{A_\delta(\omega)}(v)$ for all $v \in R$, where

$$A_\delta(\omega) = \{v \in R\colon |\mathbf{a}(v) \cdot \omega| \leq \delta\} \quad \text{with } \omega \in S^{n-1},$$

assumption (2.4) on the nondegeneracy of the velocity field yields a bound $C\delta^{\alpha/\rho_1'}$ uniformly in ω for the last term on the RHS of (4.32). Now we integrate over the sphere S^{n-1}. Recall that the Radon transform is defined as an integral over hyperplanes. Therefore, we may write

$$|\mathbf{R}\Phi_\nu(\omega, r)| = \left| \int_{\omega \cdot x = r} \Phi_\nu(x)\, dS(x) \right| \leq \int_{\omega \cdot x = r} |\Phi_\nu(x)|\, dS(x) = \mathbf{R}|\Phi_\nu|(\omega, r)$$

for $(\omega, r) \in S^{n-1} \times \mathbb{R}$. That is, the Radon transform of some function can be estimated in absolute value by the Radon transform of the absolute value of that function. Now we use the following result (cf. Ramm and Katsevich [19, Lemma 2.1.1]):

$$\int_{S^{n-1}} \mathbf{R}|\Phi_\nu|(\omega, x \cdot \omega)\, d\omega = \frac{2\pi^{\frac{n-1}{2}}}{\Gamma(\frac{n-1}{2})} \int_{\mathbb{R}^n} \frac{|\Phi_\nu(y)|}{|x-y|}\, dy$$

for all $x \in \mathbb{R}^n$. The operator on the RHS is a fractional integration of order $n-1$. We refer to Stein [20, Kapitel VIII/4.2]. Finally, we do the integration in x and use the Hölder inequality again. Then we can estimate

$$\left|\left\langle \bar{f}_{\mathbf{s},\nu}, \phi \right\rangle\right| \leq C\delta^{\alpha/\rho_1'} \|\psi\|_F \|f_\nu\|_{L_p(\mathbb{R}^n, E_1)} \left\| \int_{\mathbb{R}^n} \frac{|\Phi_\nu(y)|}{|\cdot - y|}\, dy \right\|_{L_{p'}(\mathbb{R}^n)}.$$

Let $1 < P < \infty$ be given with $\frac{1}{P} = \frac{1}{p} - \frac{n-1}{n}$. Then (cf. Stein [20])

$$\left\| \int_{\mathbb{R}^n} \frac{|\Phi_\nu(y)|}{|\cdot - y|}\, dy \right\|_{L_{p'}(\mathbb{R}^n)} \leq C \|\Phi_\nu\|_{L_{P'}(\mathbb{R}^n)}.$$

The constant does not depend on Φ_ν. Since $|\cdot|^{n-1}\rho_\nu = 2^{(\nu-1)(n-1)}[|\cdot|^{n-1}\rho_1](2^{-\nu+1}\cdot)$,

$$\|\Phi_\nu\|_{L_{P'}(\mathbb{R}^n)} \leq C 2^{(\nu-1)(n-1)} \|\phi\|_{L_{P'}(\mathbb{R}^n)}$$

for $\nu \geq 1$ with $C = \|\mathbf{F}[|\cdot|^{n-1}\rho_1]\|_{L_1(\mathbb{R}^n)}$ (the Young inequality). This proves absolute integrability for (4.31). We test $\bar{f}_{\mathbf{s},\nu}$ against all $\phi \in \mathcal{S}(\mathbb{R}^n)$ and use the fact that the Schwartz class is dense in $L_{P'}(\mathbb{R}^n)$. We obtain the following estimate:

$$\begin{aligned} \|\bar{f}_{\mathbf{s},\nu}\|_{L_P(\mathbb{R}^n)} &= \sup_{\phi \in \mathcal{S}(\mathbb{R}^n)} \|\phi\|_{L_{P'}(\mathbb{R}^n)}^{-1} \left|\left\langle \bar{f}_{\mathbf{s},\nu}, \phi \right\rangle\right| \\ &\leq C 2^{\nu(n-1)} \delta^{\alpha/\rho_1'} \|\psi\|_F \|f_\nu\|_{L_p(\mathbb{R}^n, E_1)}. \end{aligned} \tag{4.33}$$

Of course, we now want to apply a similar argument to g_ν. As above, we obtain

$$\left\langle \chi_{\mathbf{r}} \hat{g}_\nu / (i\mathbf{a} \cdot \xi), \check{\phi} \right\rangle = 2^{-\nu+1} \int_{\mathbb{R}^n} \mathbf{F}[\bar{\chi}_{\mathbf{r}} \bar{\rho}_\nu \check{\phi}](x) \cdot g_\nu(x)\, dx \quad \forall \varphi \in \mathcal{S}(\mathbb{R}^n).$$

Here, $\bar{\rho}_1(\xi) = \rho_1(\xi)/|\xi|$ and $\bar{\rho}_\nu(\xi) = \bar{\rho}_1(2^{-\nu+1}\xi)$ for all $\xi \in \mathbb{R}^n$ and $\nu \in \mathbb{N}$. But now we have a problem: for v fixed the function $\bar{\chi}_{\mathbf{r}}$ is homogeneous but not even in ξ. To circumvent this difficulty we use the Riesz transforms $\mathcal{R}_j$ with $j = 1, \ldots, n$. These mappings are defined, e.g., for all $\phi \in \mathcal{S}(\mathbb{R}^n)$ s.t. $\operatorname{supp} \check{\phi} \subset \mathbb{R}^n \backslash B_1(0)$ by

$$\mathbf{F}^{-1}[\mathcal{R}_j \phi](\xi) = i\xi_j |\xi|^{-1} \check{\phi}(\xi) \quad \forall \xi \in \mathbb{R}^n.$$

Applying these operators twice gives the following identity: $\phi = -\sum_{j=1}^n \mathcal{R}_j^2\phi$. The $\mathcal{R}_j$ are $L_p(\mathbb{R}^n)$-continuous for $1 < p < \infty$. We refer again to Stein [20] for more information and proofs. Let us define now $\Phi_{\nu,j} \in \mathcal{S}(\mathbb{R}^n)$ and $\bar{\chi}_{\mathbf{r},j}$ as follows:

$$\check{\Phi}_{\nu,j}(\xi) = i\xi_j|\xi|^{n-2}\bar{\rho}_\nu(\xi)\check{\phi}(\xi) \qquad \text{and} \qquad \bar{\chi}_{\mathbf{r},j}(\xi,v) = i\xi_j|\xi|^{-1}\bar{\chi}_{\mathbf{r}}(\xi,v)$$

for all $(\xi,v) \in \mathbb{R}^n \times R$, $j = 1,\dots,n$, and $\nu \geq 1$. Then we obtain

$$\langle \bar{f}_{\mathbf{r},\nu}, \phi\rangle = C\sum_{j=1}^n \int_R\int_{\mathbb{R}^n}\left(\int_{S^{n-1}} \bar{\chi}_{\mathbf{r},j}(\omega,v)\tilde{\Phi}_{\nu,j}(\omega, x\cdot\omega)\,d\omega\right) g_\nu(x,v)\psi(v)\,dx\,d\mu(v).$$

Note that $\bar{\chi}_{\mathbf{r},j}$ for $v \in R$ fixed is an even function in ξ. Again, we do the integration in v first and obtain for all $(x,\omega) \in \mathbb{R}^n \times S^{n-1}$ the following estimate:

$$\int_R |\psi(v)g_\nu(x,v)\bar{\chi}_{\mathbf{r}}(\omega,v)|\,d\mu(v) \leq \|\psi\|_F\|g_\nu(x)\|_{E_2}\left(\int_R \frac{\mathbf{1}_{R\setminus A_{\delta/2}(\omega)}(v)}{|i\mathbf{a}(v)\cdot\omega|^{\rho_2'}}d\mu(v)\right)^{1/\rho_2'}.$$

Using the nondegeneracy of the velocity field and Lemma 4.5 again, it is easy to show that the last factor is bounded by $C\delta^{-1+\alpha/\rho_2'}$ uniformly in ω. Proceeding as we did above, we find a constant C s.t.

$$\|\bar{f}_{\mathbf{r},\nu}\|_{L_P(\mathbb{R}^n)} \leq C2^{\nu(n-2)}\delta^{-1+\alpha/\rho_2'}\|\psi\|_F\|g_\nu\|_{L_p(\mathbb{R}^n,E_2)} \tag{4.34}$$

for $\nu \geq 1$. We used $\|\mathcal{R}_j\phi\|_{L_{P'}(\mathbb{R}^n)} \leq C\|\phi\|_{L_{P'}(\mathbb{R}^n)}$ for all $\phi \in \mathcal{S}(\mathbb{R}^n)$ and j.

5. Proofs—compactness.

First we must introduce some terminology.

5.1. Rearrangement-invariant Banach function spaces.

The Lebesgue spaces $L_p(R,\mu)$ over the finite measure space (R,μ) are just instances of more general so-called rearrangement-invariant Banach function spaces. We collect here some basic facts about these spaces but refer to Bennett and Sharpley [4] for more details.

Consider the vector space $\mathcal{M}(R,\mu)$ of μ-measurable mappings from R into $\mathbb{R}$ (or $\mathbb{C}$), where, as usual, functions which coincide μ-a.e. are identified. Denote by $\mathcal{M}^+(R,\mu)$ the set of all $f \in \mathcal{M}(R,\mu)$ with $f \geq 0$, and for some μ-measurable subset $A \subset R$ let $\mathbf{1}_A$ be its characteristic function.

Definition 5.1. *A mapping $\rho\colon \mathcal{M}^+(R,\mu) \to [0,\infty]$ is called a Banach function norm if for all $f, f_n \in \mathcal{M}^+(R,\mu)$, $a \geq 0$, and μ-measurable sets $A \subset R$ we have*

(P1) $\rho(f) = 0 \Leftrightarrow f = 0$ *μ-a.e.,*
$\rho(af) = a\rho(f)$, *and*
$\rho(f_1 + f_2) \leq \rho(f_1) + \rho(f_2)$;
(P2) $0 \leq f_1 \leq f_2$ *μ-a.e.* $\Rightarrow \rho(f_1) \leq \rho(f_2)$;
(P3) $0 \leq f_n \nearrow f_0$ *μ-a.e.* $\Rightarrow \rho(f_n) \nearrow \rho(f_0)$;
(P4) $\mu(A) < \infty \Rightarrow \rho(\mathbf{1}_A) < \infty$;
(P5) $\mu(A) < \infty \Rightarrow \int_A f\,d\mu < C\rho(f)$.

Here $C = C(\rho, A)$ is a constant not depending on f.

Definition 5.2. *Let ρ be a Banach function norm. Then the Banach function space $E = E(\rho)$ is defined as the space of all $f \in \mathcal{M}(R,\mu)$ s.t. $\|f\|_E = \rho(|f|) < \infty$.*

For each Banach function norm ρ we can define an associated norm ρ' via

$$\rho'(g) = \sup\left\{\int_R fg\,d\mu \colon f \in \mathcal{M}^+(R,\mu), \rho(f) \leq 1\right\}$$

for all $g \in \mathcal{M}^+(R,\mu)$. The associated norm ρ' also has all the properties (P1)–(P5) in Definition 5.1 and therefore generates a Banach function space $E' = E(\rho')$ associated to E. The generalized Hölder inequality holds.

Theorem 5.3. *Let E be a Banach function space, and let E' be the space associated to E. If $f \in E$ and $g \in E'$, then the product fg is absolutely integrable, and we have*

$$\int_R |fg|\, d\mu \leq \|f\|_E \|g\|_{E'}. \tag{5.1}$$

Definition 5.4. *The distribution function μ_f of $f \in \mathcal{M}(R,\mu)$ is defined by*

$$\mu_f(\lambda) = \mu\,\{x \in R\colon |f(x)| > \lambda\} \quad \forall \lambda \geq 0.$$

Note that μ_f depends only on $|f|$. If a second finite measure space (S,ν) is given, then two functions $f \in \mathcal{M}(R,\mu)$ and $g \in \mathcal{M}(S,\nu)$ are called equimeasurable if their distribution functions are identical, i.e., if $\mu_f(\lambda) = \nu_g(\lambda)$ for all $\lambda \geq 0$.

Definition 5.5. *For $f \in \mathcal{M}(R,\mu)$ let f^* be the function on $[0,\infty)$ defined by*

$$f^*(t) = \inf\,\{\lambda\colon \mu_f(\lambda) \leq t\} \quad \forall t \geq 0.$$

f^* is called the decreasing rearrangement of f.

In other words, f^* is a decreasing function on $[0,\infty)$ with the same distribution function as f itself. Here we use the convention $\inf \emptyset = \infty$; i.e., if $\mu_f(\lambda) > t$ for all $\lambda \geq 0$, then $f^*(t) = \infty$. For a finite measure space (R,μ) the distribution function μ_f is always bounded. Then f^* is a function on the interval $[0,\mu(R)]$.

Definition 5.6. *Let ρ be a Banach function norm over some finite measure space (R,μ). Then ρ is called rearrangement-invariant if $\rho(f) = \rho(g)$ for all pairs of equimeasurable functions $f,g \in \mathcal{M}(R,\mu)$. In that case we also call the Banach function space $E = E(\rho)$ defined by ρ rearrangement-invariant.*

Let us now assume that the finite measure space (R,μ) is also nonatomic, i.e., no single point carries a positive measure. This excludes Dirac measures.

Definition 5.7. *Let E be a rearrangement-invariant Banach function space over some finite nonatomic measure space (R,μ). For every t inside the interval $[0,\mu(R)]$ let A be a μ-measurable subset of R with $\mu(A) = t$. Then the function*

$$\varphi_E\colon t \mapsto \|\mathbf{1}_A\|_E \tag{5.2}$$

is called the fundamental function of E.

Since for every $B \subset R$ with $\mu(B) = t$ the functions $\mathbf{1}_A$ and $\mathbf{1}_B$ are equimeasurable, and since E is assumed rearrangement-invariant, φ_E is well defined.

The fundamental function φ_E of a rearrangement-invariant Banach function space E is quasi-concave, i.e., φ_E is increasing, $\varphi_E(t) = 0$ iff $t = 0$, and $\varphi_E(t)/t$ is decreasing. From the quasi concavity of the function, continuity in $(0,\mu(R)]$ follows. Nevertheless, a discontinuity at zero is still possible.

Theorem 5.8. *Let E be a rearrangement-invariant Banach function space over the finite nonatomic measure space (R,μ), and let E' be the associated space. Then*

$$\varphi_E(t)\varphi_{E'}(t) = t \quad \forall t \in [0,\mu(R)].$$

We refer to section II.5 in [4]. Let us discuss two examples. The fundamental function of the Lebesgue space $L^p(R,\mu)$ with $1 \leq p < \infty$ is given by $\varphi_{L^p}(t) = t^{1/p}$ for

$t \in [0, \mu(R)]$. However, if $p = \infty$, then

$$\varphi_{L^\infty}(t) = \begin{cases} 0 & \text{for } t = 0, \\ 1 & \text{for } t \in (0, \mu(R)], \end{cases}$$

since the characteristic function of a null set is equivalent to the zero function. In that case, the fundamental function is discontinuous at zero.

In applications, the Orlicz space $L \log L(R, \mu)$ plays an important role. That is the rearrangement-invariant Banach function space defined by the norm

$$\|f\|_{L \log L(R,\mu)} = -\int_0^{\mu(R)} f^*(t) \log\left(t/\mu(R)\right)\, dt.$$

For the fundamental function we have

$$\varphi_{L \log L}(t) = t\left(1 - \log\left(t/\mu(R)\right)\right) \quad \forall t \in [0, \mu(R)].$$

The space associated to $L \log L(R, \mu)$ is the space $\exp L(R, \mu)$ of exponentially integrable functions. The corresponding fundamental function is given by

$$\varphi_{\exp L}(t) = \frac{1}{1 - \log\left(t/\mu(R)\right)} \quad \forall t \in [0, \mu(R)].$$

Note that $\varphi_{\exp L}$ is continuous at zero (but not differentiable). That is remarkable because for all $1 < p < \infty$ the following inclusions hold (since $\mu(R) < \infty$):

$$L^\infty(R, \mu) \hookrightarrow \exp L(R, \mu) \hookrightarrow L^p(R, \mu) \hookrightarrow L \log L(R, \mu) \hookrightarrow L^1(R, \mu).$$

Although $\exp L$ and L^∞ are so close that no other L^p-space fits between them, there is a considerable difference in their respective fundamental functions.

5.2. Proof of Theorem 2.5. Again, we briefly recall our assumptions. Let E_1 and E_2 be two rearrangement-invariant Banach function spaces over some nonatomic finite measure space. Assume we are given a sequence of pairs

$$f^{(k)} \in B^s_{p,q}(\mathbb{R}^n, E_1) \quad \text{and} \quad g^{(k)} \in B^{s-\tau}_{p,q}(\mathbb{R}^n, E_2)$$

satisfying the transport equation (2.1) in $\mathcal{S}'(\mathbb{R}^n, E)$ for $E = E_1 + E_2$. We fix a weight ψ in some subset $F \subset E'$ s.t. multiplication with ψ maps E_1 continuously into some rearrangement-invariant Banach function space G_1 and, similarly, E_2 into some G_2. If the velocity field now satisfies a nondegeneracy condition and if the fundamental function of G_1' is continuous at zero, we will show that the sequence of averages $\bar{f}^{(k)}$ is precompact in a suitable local Besov space $B^{S,\mathrm{loc}}_{P,q}(\mathbb{R}^n)$.

To simplify notation, we will drop the index k in the following. We already know that there exists a decomposition of the average $\bar{f}$ into $\bar{f} = \sum_{\nu=0}^\infty \bar{f}_\nu$ in $\mathcal{S}'(\mathbb{R}^n)$ and $\bar{f}_\nu = \bar{f}_{\mathbf{s},\nu} + \bar{f}_{\mathbf{r},\nu}$ for $\nu \geq 1$. In contrast to our approach in section 4, we will now choose the splitting parameter δ independent of ν. Then we can write $\bar{f} = \bar{f}_0 + \bar{F}_{\mathbf{s}} + \bar{F}_{\mathbf{r}}$ with $\bar{F}_{\mathbf{s}} = \sum_{\nu=1}^\infty \bar{f}_{\mathbf{s},\nu}$ and $\bar{F}_{\mathbf{r}} = \sum_{\nu=1}^\infty \bar{f}_{\mathbf{r},\nu}$. We claim that $\bar{f}_0 \in B^\sigma_{P,q}(\mathbb{R}^n)$ for arbitrary $\sigma \in \mathbb{R}$ and P as given above. Moreover,

$$\bar{F}_{\mathbf{s}} \in B^S_{P,q}(\mathbb{R}^n) \quad \text{and} \quad \bar{F}_{\mathbf{r}} \in B^{S+\epsilon}_{P,q}(\mathbb{R}^n)$$

with the S from above and $\epsilon = 1 - \tau > 0$. To see that, let us first consider the dyadic components $\bar{F}_{\mathbf{s},\nu}$. As an immediate consequence of the support properties of the family φ_ν with $\nu \geq 0$, we realize that

$$\bar{F}_{\mathbf{s},\nu} = \mathbf{F}^{-1}[\varphi_\nu \mathbf{F}\bar{F}_{\mathbf{s}}] = \begin{cases} \mathbf{F}^{-1}[\varphi_0 \mathbf{F}\bar{f}_{\mathbf{s},1}] & \text{if } \nu = 0, \\ \mathbf{F}^{-1}[\varphi_1 \mathbf{F}(\bar{f}_{\mathbf{s},1} + \bar{f}_{\mathbf{s},2})] & \text{if } \nu = 1, \\ \bar{f}_{\mathbf{s},\nu} & \forall \nu \geq 2. \end{cases}$$

Since $P \geq 1$, we can now apply the Young inequality to obtain the estimates

$$\begin{aligned} &\|\bar{F}_{\mathbf{s},0}\|_{L_P(\mathbb{R}^n)} \leq C\|\bar{f}_{\mathbf{s},1}\|_{L_P(\mathbb{R}^n)}, \\ &\|\bar{F}_{\mathbf{s},1}\|_{L_P(\mathbb{R}^n)} \leq C\left\{\|\bar{f}_{\mathbf{s},1}\|_{L_P(\mathbb{R}^n)} + \|\bar{f}_{\mathbf{s},2}\|_{L_P(\mathbb{R}^n)}\right\}, \quad \text{and} \\ &\|\bar{F}_{\mathbf{s},\nu}\|_{L_P(\mathbb{R}^n)} = \|\bar{f}_{\mathbf{s},\nu}\|_{L_P(\mathbb{R}^n)} \quad \forall \nu \geq 2. \end{aligned}$$

However, the L_P-norm of $\bar{f}_{\mathbf{s},\nu}$ has already been estimated in the last sections. Using the generalized Hölder inequality for the Banach function space G_1, we obtain the following analogue of the estimates (4.25) and (4.33):

$$\|\bar{f}_{\mathbf{s},\nu}\|_{L_P(\mathbb{R}^n)} \leq C2^{\nu(\frac{1}{p}-\frac{1}{P})} \sup_{\xi \in \mathbb{R}^n} \|\mathbf{1}_{A_\delta(\xi)}\|_{G_1'} \|\psi\|_F \|f_\nu\|_{L_p(\mathbb{R}^n,E_1)}.$$

The constant does not depend on f or ν. But the G_1'-norm of the characteristic function of some set A with measure $s \geq 0$ is just the fundamental function $\varphi_{G_1'}(s)$. And since this function is increasing, we can use the nondegeneracy condition for $\mathbf{a}$ to obtain the following estimate (cf. (2.9)):

$$\|\bar{F}_{\mathbf{s}}\|_{B^S_{P,q}(\mathbb{R}^n)} \leq C\varphi_{G_1'}(\eta(\delta))\,\|\psi\|_F \|f\|_{B^s_{p,q}(\mathbb{R}^n,E_1)}. \tag{5.3}$$

Again, the constant C does not depend on f or δ. In the same way, we can find a bound for the second term $\bar{F}_{\mathbf{r}}$. We use the generalized Hölder inequality for G_2, estimate $|i\mathbf{a}\cdot\omega|^{-1}\mathbf{1}_{R\setminus A_{\delta/2}(\omega)}$ by $C\delta^{-1}$, and obtain

$$\|\bar{F}_{\mathbf{r}}\|_{B^{S+\epsilon}_{P,q}(\mathbb{R}^n)} \leq C\delta^{-1}\|\psi\|_F \|g\|_{B^{s-\tau}_{p,q}(\mathbb{R}^n,E_2)}. \tag{5.4}$$

Finally, note that most dyadic parts of $\bar{f}_0$ vanish by the construction of φ_ν. Using the Young, the Nikol'skij, and the Hölder inequalities (3.4)–(5.1), we then obtain

$$\|\bar{f}_0\|_{B^\sigma_{P,q}(\mathbb{R}^n)} \leq C\|\psi\|_F \|f\|_{B^s_{p,q}(\mathbb{R}^n,E_1)} \quad \forall \sigma \in \mathbb{R}.$$

Let us now consider sequences $f^{(k)}, g^{(k)}$ with (2.8), satisfying a transport equation (2.1). Let $\chi \in \mathcal{D}(\mathbb{R}^n)$ be some test function with compact support. Then we may decompose $\chi\bar{f}^{(k)}$ into three parts, $\chi\bar{F}_{\mathbf{s}}^{(k)}, \chi\bar{F}_{\mathbf{r}}^{(k)}$, and $\chi\bar{f}_0^{(k)}$, and estimate as follows.

I. By assumption, the fundamental function of the rearrangement-invariant Banach function space G_1' is continuous at zero, and $\lim_{\delta\to 0}\eta(\delta) = 0$. Therefore,

$$\|\chi\bar{F}_{\mathbf{s}}^{(k)}\|_{B^S_{P,q}(\mathbb{R}^n)} \leq C\|\bar{F}_{\mathbf{s}}^{(k)}\|_{B^S_{P,q}(\mathbb{R}^n)} \leq C\varphi_{G_1'}(\eta(\delta))\,\|\psi\|_F \|f^{(k)}\|_{B^s_{p,q}(\mathbb{R}^n,E_1)} \longrightarrow 0$$

uniformly in k if $\delta \to 0$: The sequence $f^{(k)}$ is uniformly bounded, and the constants are independent of k and δ. We also used Theorem 3.12 and (5.3). We conclude that the first part of $\bar{f}^{(k)}$ becomes small, choosing the parameter δ suitably.

II. Because of Theorem 3.12 and (5.4), there exists a number $C > 0$ s.t.

$$\|\chi\bar{F}_{\mathbf{r}}^{(k)}\|_{B^{S+\epsilon}_{P,q}(\mathbb{R}^n)} \leq C\|\bar{F}_{\mathbf{r}}^{(k)}\|_{B^{S+\epsilon}_{P,q}(\mathbb{R}^n)} \leq C\delta^{-1}\|\psi\|_F\|g^{(k)}\|_{B^{s-\tau}_{p,q}(\mathbb{R}^n,E_2)} \tag{5.5}$$

with C independent of δ, k and $g^{(k)}$. Therefore, for every fixed $\delta > 0$, the sequence $\chi\bar{F}_{\mathbf{r}}^{(k)}$ is uniformly bounded in $B^{S+\epsilon}_{P,q}(\mathbb{R}^n)$ for some $\epsilon > 0$. And since the supports of all functions of the sequence are contained in one single compact subset of $\mathbb{R}^n$, we can use, e.g., Theorem 3.3.2/1 in Edmunds and Triebel [11] to conclude that $\chi\bar{F}_{\mathbf{r}}^{(k)}$ is precompactly contained in $B^S_{P,q}(\mathbb{R}^n)$. Similarly, we proceed for $\chi\bar{f}_0^{(k)}$.

Therefore, the sequence of averages $\chi\bar{f}^{(k)}$ can be decomposed into three parts, two of which are precompact in $B^S_{P,q}(\mathbb{R}^n)$, while the third goes to zero uniformly with respect to k, as $\delta \to 0$. Hence $\chi\bar{f}^{(k)}$ itself is precompact as claimed. The proof is complete.

Remark 5.9. Note that the fine structure of G_2 plays no role in the proof of Theorem 2.5. If $\epsilon = 0$, precompactness of $\chi\bar{F}_{\mathbf{r}}^{(k)}$ in $B^S_{P,q}(\mathbb{R}^n)$ follows from estimate (5.5) and the assumed precompactness of $g^{(k)}$ in $B^{s-1}_{p,q}(\mathbb{R}^n, E_2)$.

Remark 5.10. We return to Remark 2.8. Assume that

$$f^{(k)} \in B^{s,\mathrm{loc}}_{p,q}(\mathbb{R}^n, E_1) \quad \text{and} \quad g^{(k)} \in B^{s-\tau,\mathrm{loc}}_{p,q}(\mathbb{R}^n, E_2)$$

are uniformly bounded; i.e., for all test functions $\chi \in \mathcal{D}(\mathbb{R}^n)$ the sequence $\chi f^{(k)}$ is bounded in $B^s_{p,q}(\mathbb{R}^n, E_1)$, etc. Then we have the following equality:

$$\mathrm{div}_x(\mathbf{a}\chi f^{(k)}) = \chi g^{(k)} - f^{(k)}\mathrm{div}_x(\mathbf{a}\chi).$$

The RHS is uniformly bounded in $B^{s-\tau}_{p,q}(\mathbb{R}^n, E)$ with $E = E_1 + E_2$. From Theorem 2.5 we therefore conclude that the sequence $\chi\bar{f}^{(k)}$ is precompact in $B^S_{P,q}(\mathbb{R}^n)$. Hence it is also possible to choose $f^{(k)}$ and $g^{(k)}$ in local Besov spaces only.

REFERENCES

[1] V. Agoshkov, *Spaces of functions with differential-difference characteristics and the smoothness of solutions of the transport equation*, Soviet Math. Dokl., 29 (1984), pp. 662–666.

[2] H. Amann, *Operator-valued Fourier multipliers, vector-valued Besov spaces, and applications*, Math. Nachr., 186 (1997), pp. 5–56.

[3] M. Bézard, *Régularité L^p précisée des moyennes dans les équations de transport*, Bull. Soc. Math. France, 122 (1994), pp. 27–76.

[4] C. Bennett and R. Sharpley, *Interpolation of Operators*, Academic Press, Boston, 1988.

[5] F. Bouchut, *Introduction à la théorie mathématique des équations cinétiques*, Session “L'Etat de la Recherche” de la S.M.F. Equations Cinétiques, 1998.

[6] F. Castella and B. Perthame, *Estimations de Strichartz pour les équations de transport cinétique*, C. R. Acad. Sci. Paris Sér. I Math., 322 (1996), pp. 535–540.

[7] R. DeVore and G. P. Petrova, *The Averaging Lemma*, preprint, University of South Carolina, Columbia, SC, 1999.

[8] R. J. DiPerna and P.-L. Lions, *On the Cauchy problem for the Boltzmann equation: Global existence and weak stability results*, Ann. of Math. (2), 130 (1989), pp. 321–366.

[9] R. J. DiPerna and P.-L. Lions, *Global weak solutions of Vlasov-Maxwell systems*, Comm. Pure Appl. Math., 42 (1989), pp. 729–757.

[10] R. J. DiPerna, P.-L. Lions, and Y. Meyer, *L^p regularity of velocity averages*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 8 (1991), pp. 271–287.

[11] D. E. Edmunds and H. Triebel, *Function Spaces, Entropy Numbers and Differential Operators*, Cambridge University Press, Cambridge, UK, 1996.

[12] P. Gérard, *Moyennisation et régularité deux-microlocale*, Ann. Sci. École Norm. Sup. (4), 23 (1990), pp. 89–121.

[13] F. Golse, P.-L. Lions, B. Perthame, and R. Sentis, *Regularity of the moments of the solution of a transport equation*, J. Funct. Anal., 76 (1988), pp. 110–125.

[14] F. Golse, B. Perthame, and R. Sentis, *Un résultat de compacité pour les équations de transport et applications au calcul de la limite de la valeur propre principale d'un opérateur de transport*, C. R. Acad. Sci. Paris Sér. I Math., 301 (1985), pp. 341–344.

[15] L. Hörmander, *The Analysis of Linear Partial Differential Operators*, Springer, Berlin, Heidelberg, 1985.

[16] P.-L. Lions, B. Perthame, and E. Tadmor, *A kinetic formulation of multidimensional scalar conservation laws and related equations*, J. Amer. Math. Soc., 7 (1994), pp. 169–191.

[17] B. Perthame, *Time decay, propagation of low moments and dispersive effects for kinetic equations*, Comm. Partial Differential Equations, 21 (1996), pp. 659–686.

[18] B. Perthame and P. E. Souganidis, *A limiting case for velocity averaging*, Ann. Sci. École Norm. Sup. (4), 31 (1998), pp. 591–598.

[19] A. G. Ramm and A. I. Katsevich, *The Radon Transform and Local Tomography*, CRC Press, Boca Raton, FL, 1996.

[20] E. M. Stein, *Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals*, Princeton University Press, Princeton, NJ, 1993.

[21] H. Triebel, *Theory of Function Spaces*, Birkhäuser, Basel, 1983.

[22] H. Triebel, *Fractals and Spectra*, Birkhäuser, Basel, 1997.

[23] M. Westdickenberg and S. Noelle, *A new convergence proof for finite volume schemes using the kinetic formulation of conservation laws*, SIAM J. Numer. Anal., 37 (2000), pp. 742–757.

SIAM J. MATH. ANAL.
Vol. 33, No. 5, pp. 1033–1047

TIME EVOLUTION OF NEMATIC LIQUID CRYSTALS WITH VARIABLE DEGREE OF ORIENTATION*

M. C. CALDERER[†], D. GOLOVATY[‡], F.-H. LIN[§], AND C. LIU[†]

Abstract. We consider an evolution system, describing the time-dependent behavior of nematic liquid crystals with variable degree of orientation within the continuum model of Ericksen. We establish a dissipation relation and prove both the global existence of weak solutions and the local existence of classical solutions. Furthermore, we investigate the stability and long-time behavior of solutions and obtain an exact solution of the corresponding stationary system in a one-dimensional case.

Key words. nematic liquid crystals, reaction-diffusion mechanism, weak solutions, free energy functional

AMS subject classification. 35Q35

PII. S0036141099362086

1. Introduction.

In this paper, we study the system of partial differential equations governing the evolution of uniaxially nematic liquid crystals with variable degree of orientation in the absence of either flow or electromagnetic field.

A typical nematic liquid crystal consists of the rigid, rodlike molecules with one molecular axis being much longer than the other two. As a consequence of the highly anisotropic molecular shape, there is a higher probability that the axes of any two neighboring molecules point in a similar direction. The resulting molecular structure is then commonly characterized as having some orientational order.

Within the standard continuum theory of Ericksen [4], the local molecular orientation is described through the second moments of the appropriate local orientational distribution functions. For a uniaxial nematic, the second moment is fully specified by a unit vector $\mathbf{n}$, called the director, and a scalar s, called the order parameter. Given a point $\mathbf{x}$, a simple interpretation of these characteristics is that, on average, the molecules tend to be oriented along the director $\mathbf{n}(\mathbf{x})$ in a vicinity of $\mathbf{x}$. Since the majority of the molecules actually point in the directions close to, but different than $\mathbf{n}(\mathbf{x})$, the quality of the intermolecular alignment has to be described by an additional parameter $s(\mathbf{x})$. It can be shown that the values of s are restricted to the interval $(-\frac{1}{2}, 1)$, where all molecules align perfectly with the director as s approaches 1 and align perpendicular to it as s approaches $-\frac{1}{2}$. The special case of $s = 0$ corresponds to the isotropic state with randomly oriented molecules in which the director $\mathbf{n}$ is meaningless. The important consequence of the isotropy for the analysis of the Ericksen model is that the underlying partial differential equations become singular when $s = 0$.

*Received by the editors September 23, 1999; accepted for publication (in revised form) August 10, 2001; published electronically January 16, 2002.

http://www.siam.org/journals/sima/33-5/36208.html

[†] Department of Mathematics and Center for Materials Physics, Penn State University, University Park, PA 16802 (mcc@math.psu.edu, liu@math.psu.edu). The first author's work was supported in part by NSF grant DMS-9704714, 1997–1999.

[‡] Department of Mathematics and Computer Science, The University of Akron, Akron, OH 44325 (dmitry@uakron.edu).

[§] Courant Institute, New York University, 251 Mercer Street, New York, NY 10012 (linf@cims.nyu.edu). This author's work was supported in part by NSF grant DMS-9896396, 1998–2001.

The two-parameter approach of Ericksen is a generalization of the earlier Oseen–Frank theory in which a liquid crystal configuration is described by the director field alone. The well-known flaw of the Oseen–Frank approach is its inability to predict some observable liquid crystal configurations. In particular, for the appropriately chosen material constants, the energy minimizing configuration $\mathbf{n}$ is a harmonic map into S^2 with a singular set consisting at most of isolated points [13]. This, however, contradicts the experimental observations [8] of both line and surface defects in various materials, including liquid crystals. The Ericksen model was introduced, in part, in order to address this deficiency of the Oseen–Frank theory.

When s is bounded away from zero and the gradient of the director can be controlled, the evolution equation for the order parameter (2.9) exhibits a reaction-diffusion mechanism normally responsible for the dynamics of the domain structures. This mechanism, however, might be suppressed when s is small and, therefore, the size of the term involving the gradient of the director is very large. The evolution and the interactions of the structural defects corresponding to vanishing s constitutes, in fact, a difficult mathematical problem, especially for the defects of codimension one or two (surface and line defects). From the practical point of view, the understanding of the defect dynamics plays an important role in the industrial applications of liquid crystals, such as the design of efficient display devices.

The main goal of this paper is to establish both the existence and the regularity results for the system of partial differential equations comprising the Ericksen model. Here, the biggest challenge is to find the framework within which a degenerate parabolic equation is guaranteed to have a solution. Our approach is to approximate the set of the governing equations by a system of nonsingular PDEs and to show that the approximate solutions converge, in a weak sense, to a solution of the Ericksen system.

Following [4] and [10], we consider a free energy functional with the density having a quadratic dependence on both $s\nabla\mathbf{n}$ and ∇s. Furthermore, we assume that the free energy density depends on the order parameter through the double-well potential $W(s)$. The locations of the extrema for W and their corresponding energy values are physically determined by the temperature in thermotropic liquid crystals and by the concentration in lyotropic liquid crystals. In particular, a minimum of W can be moved into the isotropic state $s = 0$ either by raising the temperature, or by lowering the concentration of a polymer in a given solvent.

In the stationary case, the model for nematic liquid crystals with variable degree of orientation was studied by Leslie [10] and by Virga [15]. Some aspects of the time dependent behavior in modeling of patterns, associated with flow relaxation, have been investigated by Calderer [1]. Recently, the plane Poiseuille flow was considered by Calderer and Liu [2] to show that the order parameter allows for the modeling of new types of defects.

The outline of the paper is as follows. The governing equations and the constitutive properties of the model are outlined in section 2. In section 3 we use the dissipative relation to obtain some a priori estimates. These relations, in turn, allow us to prove the global existence of weak solutions in section 4. Next, using the semigroup methods, we prove the local in time existence of classical solutions (section 5) and then, in section 6, discuss both the stability and the long time behavior of solutions. Finally, an exact solution of the corresponding stationary system is obtained in section 7 in order to illustrate the singular behavior of the Ericksen system in the one-dimensional case.

2. Governing equations. In the absence of a flow, the evolution for s and $\mathbf{n}$ is governed by the following system of partial differential equations [4]:

$$\beta(s)\dot{s} = \operatorname{div}\left\{\frac{\partial W_1}{\partial \nabla s}\right\} - \frac{\partial W_1}{\partial s} - W'(s), \tag{2.1}$$

$$\gamma(s)\dot{\mathbf{n}} \times \mathbf{n} = \left(\operatorname{div}\left\{\frac{\partial W_1}{\partial \nabla \mathbf{n}}\right\} - \frac{\partial W_1}{\partial \mathbf{n}}\right) \times \mathbf{n}, \tag{2.2}$$

$$|\mathbf{n}|^2 = 1, \tag{2.3}$$

where $\mathcal{W}(s, \mathbf{n}, \nabla s, \nabla \mathbf{n}) = W(s) + W_1(s, \mathbf{n}, \nabla s, \nabla \mathbf{n})$ denotes the Helmholtz free energy. As in the Oseen–Frank theory [16], the term W_1 is a quadratic function of the gradients of $\mathbf{n}$ and s. The absence of the first order term in free energy is due to the invariance of $\mathcal{W}$ with respect to the transformation $\mathbf{n} \to -\mathbf{n}$. From now on we will assume that

(2.4) $W_1 = \dfrac{1}{2}k_1|\nabla s|^2 + \dfrac{1}{2}k_2 s^2 |\nabla \mathbf{n}|^2$, where k_1 and k_2 are positive constants,

(2.5) $W(s)$ is a smooth double-well potential satisfying $\lim\limits_{s \to -\frac{1}{2}, 1} W(s) = \infty$.

Using the molecular theory predictions [4] on the asymptotic behavior of the constitutive functions near $s = 0$ and rescaling we set

$$\gamma(s) = \frac{k_2}{k_3} s^2, \tag{2.6}$$

$$\beta(s) \equiv 1, \tag{2.7}$$

where k_3 is a positive constant.

Suppose that the material occupies a bounded domain $\Omega \in \mathbf{R}^n$ with the smooth boundary $\partial\Omega$, where $n \leq 3$. We denote the points in Ω by $x = (x_1, \ldots, x_n)$ and consider the planar director configurations

$$\mathbf{n} = (\cos\phi, \sin\phi, 0). \tag{2.8}$$

Then (2.1)–(2.2) reduce to

$$s_t = k_1 \triangle s - k_2 s |\nabla \phi|^2 - W'(s), \tag{2.9}$$

$$s^2 \phi_t = k_3 \operatorname{div}\left(s^2 \nabla \phi\right), \tag{2.10}$$

where $x \in \Omega$ and $t > 0$. This system will be supplemented by the following initial and boundary data:

$$s(x,0) = g(x), \quad \phi(x,0) = h(x), \qquad x \in \Omega, \tag{2.11}$$

$$s(x,t) = g(x), \quad \phi(x,t) = h(x), \qquad x \in \partial\Omega, \quad t \geq 0, \tag{2.12}$$

where g and h are the given smooth functions of x.

Remark 2.1. It has been shown by several authors [11], [15] that, in the stationary case, the introduction of s into the energy allows for the modeling of new types of defects. Consider, for example, the Oseen–Frank free energy for $\mathbf{n}(x)$ with $x \in \mathbf{R}$. Then the corresponding Euler–Lagrange equation is

$$-\mathbf{n}'' = |\mathbf{n}'|^2 \mathbf{n}.$$

Multiplying this equation by $\mathbf{n}'$ and using the fact that $|\mathbf{n}| = 1$, we have

$$(|\mathbf{n}'|^2)' = 0, \text{ and } -\mathbf{n}'' = a\mathbf{n},$$

where $a \geq 0$ is a constant. Therefore, $\mathbf{n}$ is a smooth, defect-free vector field. On the other hand, when the order parameter and the corresponding energy contribution are included into the model, the defects can exist in one-dimensional geometries [2].

3. A priori estimates and the Lyapunov function.

In this section we construct a Lyapunov function and obtain some a priori estimates on the solutions of (2.9)–(2.10).

THEOREM 3.1. *Suppose that the pair* $(s(x\,,\,t),\,\phi(x\,,\,t))$ *is a classical solution of the system* (2.9)–(2.10) *for* $x \in \Omega$ *and* $t \in (0\,,\,T)$. *Set*

$$E(t) := E[s(\cdot,t),\,\phi(\cdot,t)] = \int_\Omega \left[\frac{k_1}{2}|\nabla s|^2 + \frac{k_2}{2}s^2|\nabla\phi|^2 + W(s)\right] dx. \tag{3.1}$$

Then $E(t)$ *is a Lyapunov function for the system* (2.9)–(2.10)*, that is,*

$$E(t) \geq m\,|\Omega| > -\infty, \tag{3.2}$$

$$E\,'(t) \leq 0 \textit{ for all } t \in (0,T), \tag{3.3}$$

where $|\Omega|$ *denotes the Lebesgue measure of* Ω *and* $m := \min_{s\in(-\frac{1}{2},1)} W(s)$.

Proof. First, we calculate

$$E\,'(t) = \int_\Omega \left[\Big(k_2 s|\nabla\phi|^2 + W'(s)\Big)\, s_t + k_1\nabla s\cdot\nabla s_t + k_2 s^2\nabla\phi\cdot\nabla\phi_t\right] dx. \tag{3.4}$$

Integrating the mixed derivatives terms in (3.4) by parts and using the boundary conditions (2.12) along with the governing equations (2.9)–(2.10), we have

$$\begin{aligned} E\,'(t) &= \int_\Omega \left\{ \Big(-k_1\Delta s + k_2 s|\nabla\phi|^2 + W'(s)\Big)\; s_t - k_2\phi_t \operatorname{div}\left(s^2\nabla\phi\right) \right\} dx \\ &= -\int_\Omega \left\{ s_t^2 + \frac{k_2}{k_3}s^2{\phi_t}^2 \right\} dx \leq 0\,. \end{aligned} \tag{3.5}$$

Therefore, $E(t) \leq E(0)$ for all $t \in [0,T]$. The proof of the assertion (3.2) follows immediately from nonnegativity of the gradient terms in (3.1). □

Note that, from the previous theorem, one can obtain the following bound on the gradient of the director:

$$\int_\Omega s^2|\nabla\phi|^2\,dx \leq 2k_2^{-1}\left\{E(0) - m\,|\Omega|\right\} \tag{3.6}$$

for all $t \in [0,T]$.

LEMMA 3.2. *Let* Ω *be a domain with the smooth boundary and* (s,ϕ) *be a smooth solution of the system* (2.9)–(2.10). *Suppose that* $g(x) \in \left(-\frac{1}{2},1\right)$ *for all* $x \in \bar\Omega$*; then there exist two constants,* s_- *and* s_+*, such that* $-\frac{1}{2} < s_- < 0 < s_+ < 1$ *and* $s(x,t) \in (s_-,\,s_+)$ *for all* $x \in \Omega$ *and* $t > 0$.

Proof. This result follows from the maximum principle [12] by taking into account the property (2.5) of the function W. By our assumptions on the function g, there exist

$$g_- := \min_{x\in\bar\Omega} g(x) \text{ and } g_+ := \max_{x\in\bar\Omega} g(x),$$

where $-\frac{1}{2} < g_- \leq g_+ < 1$. Moreover, since W' is continuous and $\lim_{s\to -\frac{1}{2},1} W(s) = \infty$, we can find

$$s_- \in (-1/2, \min\{0, g_-\}),$$

such that $W'(s_-) < 0$, and

$$s_+ \in (\max\{0, g_+\}, 1),$$

such that $W'(s_+) > 0$.

Next, suppose that $t^* > 0$ is the smallest time for which $s(x^*, t^*) = s_+$ for some $x^* \in \Omega$. Then $W'(s(x^*, t^*)) = W'(s_+) > 0$, $s_t(x^*, t^*) \geq 0$, and $\triangle s(x^*, t^*) \leq 0$. This, however, contradicts (2.9); hence $s(x, t) < s_+$ for all $x \in \Omega$ and $t > 0$. A similar argument can be used to show that $s > s_-$ as well. □

By slightly modifying the proof of the previous lemma one can also prove the following lemma.

LEMMA 3.3. *Suppose that s_1 and s_2 are the wells of the potential W; that is, both s_1 and s_2 are the strict local minima of W. Let Ω be a domain with the smooth boundary and (s, ϕ) be the smooth solution of the system* (2.9)–(2.10). *Suppose that $g(x) \in (s_1, s_2)$ for all $x \in \bar{\Omega}$; then $s(x, t) \in (s_1, s_2)$ for all $x \in \Omega$ and $t > 0$.*

In the next two sections we discuss the solvability of the problem (2.9)–(2.10).

4. Global existence of weak solutions. The system (2.9)–(2.10) for the nematic liquid crystals with variable degree of orientation was originally introduced by Ericksen to model the line singularities. However, the obvious drawback of (2.9)–(2.10) is that the problem loses its parabolicity whenever the order parameter s is allowed to vanish. In order to overcome the degeneracy in the stationary version of (2.9)–(2.10), Lin [11] introduced a variable $\mathbf{u} = s\mathbf{n}$ and recast the problem in terms of s and $\mathbf{u}$. Within the new formulation, the constraint $|\mathbf{n}|^2 = 1$ has to be replaced by $s^2 = |\mathbf{u}|^2$, and the flow for $(s, \mathbf{u})$ is from Ω into a circular cone. The corresponding problem is then nonsingular, but since the target manifold itself is not smooth, the solution of the reformulated problem is, in general, only Lipschitz continuous.

Here, in order to overcome the degeneracy in (2.10) at $s = 0$, we introduce the modified system of equations

$$s_{\epsilon t} = k_1 \triangle s_\epsilon - k_2 s_\epsilon |\nabla \phi_\epsilon|^2 - W'(s_\epsilon), \tag{4.1}$$

$$\left(s_\epsilon^2 + \epsilon^2\right) \phi_{\epsilon t} = k_3 \operatorname{div}\left(\left(s_\epsilon^2 + \epsilon^2\right) \nabla \phi_\epsilon\right), \tag{4.2}$$

where $\epsilon > 0$ is small. Suppose that $(s_\epsilon, \phi_\epsilon)$ satisfy the initial and boundary data (2.11)–(2.12). Then the proof of the following lemma is the same as that of Theorem 3.1.

LEMMA 4.1. *Suppose that $(s_\epsilon, \phi_\epsilon)$ is a smooth solution of* (4.1)–(4.2). *Then it satisfies the following dissipative relation:*

$$E'_\epsilon(t) = -\int_\Omega \left[\left(s_{\epsilon t}^2 + \frac{k_2}{k_3}\left(s_\epsilon^2 + \epsilon^2\right)\phi_{\epsilon t}^2\right)\right] dx \leq 0, \tag{4.3}$$

where

$$E_\epsilon(t) = \int_\Omega \left[\frac{k_1}{2}|\nabla s_\epsilon|^2 + \frac{k_2}{2}\left(s_\epsilon^2 + \epsilon^2\right)|\nabla \phi_\epsilon|^2 + W(s_\epsilon)\right] dx. \tag{4.4}$$

Theorem 4.2. *There exists a unique global classical solution* $(s_\epsilon, \phi_\epsilon)$ *of the system* (4.1)–(4.2)*, corresponding to the initial and boundary data* (2.11)–(2.12).

Here the existence follows from the standard results for parabolic systems, along with a priori estimates derived in Lemma 4.1 [9, Chapter 10], [5, Chapter 3]. The regularity of solutions follows from Sobolev imbedding result, combined with a bootstrap argument.

Note that, for any $\epsilon > 0$, the conclusion of the Lemma 3.2 can be applied to the solutions of (4.1)–(4.2), since (4.1) for s_ϵ is the same as (2.9) for s. Then s_ϵ is bounded in $L^\infty(\Omega_T)$ uniformly in ϵ, where $\Omega_T = \Omega \times (0, T)$. Moreover, for any $\epsilon > 0$ and $0 < T < \infty$, the solutions of (4.1)–(4.2) belong to the following functional spaces:

$$\begin{gathered} s_\epsilon \in L^\infty\left(0,T;H^1(\Omega)\right) \cap H^1\left(0,T;L^2(\Omega)\right), \\ s_\epsilon \nabla\phi_\epsilon \in L^\infty\left(0,T;L^2(\Omega)\right), \\ s_\epsilon \phi_{\epsilon t} \in L^2\left(0,T;L^2(\Omega)\right). \end{gathered}$$

In particular, s_ϵ is uniformly bounded with respect to ϵ in $L^\infty\left(0,T;H^1(\Omega)\right)$, while $s_\epsilon\nabla\phi_\epsilon$ is uniformly bounded with respect to ϵ in $L^\infty\left(0,T;L^2(\Omega)\right)$, and $s_{\epsilon t}$ is uniformly bounded with respect to ϵ in $L^2\left(0,T;L^2(\Omega)\right)$. Furthermore, ϕ_ϵ is uniformly bounded with respect to ϵ in $L^\infty(\Omega_T)$ by the standard version of the maximum principle [12] for (4.2) and the assumptions on h in (2.11)–(2.12).

Theorem 4.3. *There exists a global weak solution,* (s, ϕ)*, of the system*

$$\frac{1}{2}\left(s^2\right)_t = \frac{k_1}{2}\triangle\left(s^2\right) - k_1|\nabla s|^2 - k_2 s^2|\nabla\phi|^2 - W'(s)\,s, \tag{4.5}$$

$$s^2\phi_t = k_3\,\mathrm{div}\left(s^2\nabla\phi\right). \tag{4.6}$$

Moreover,

$$\begin{gathered} s \in L^\infty\left(0,T;H^1(\Omega)\right) \cap H^1\left(0,T;L^2(\Omega)\right), \\ s\nabla\phi \in L^\infty\left(0,T;L^2(\Omega)\right), \\ s,\, \phi \in L^\infty(\Omega_T). \end{gathered}$$

Proof. We multiply (4.1) and (4.2) by smooth test functions, $s_\epsilon\xi_1(t)\xi_2(x)$ and $\eta_1(t)\eta_2(x)$, respectively, and pass to the limit as $\epsilon \to 0$. First, we carry out several integrations by parts:

$$\begin{aligned} &-\int_0^T\int_\Omega \frac{1}{2}s_\epsilon^2\xi_{1t}(t)\,\xi_2(x)\,dx\,dt \\ &= \int_0^T\int_\Omega \Big\{ \frac{k_1}{2}s_\epsilon^2\xi_1(t)\triangle\xi_2(x) - k_1|\nabla s_\epsilon|^2\xi_1(t)\,\xi_2(x) \\ &\qquad - k_2 s_\epsilon^2|\nabla\phi_\epsilon|^2\xi_1(t)\,\xi_2(x) - W'(s_\epsilon)\,s_\epsilon\xi_1(t)\,\xi_2(x) \Big\}\,dx\,dt \end{aligned} \tag{4.7}$$

and

$$\begin{aligned} &-\int_0^T\int_\Omega \left\{ \left(s_\epsilon^2+\epsilon^2\right)\phi_\epsilon\eta_{1t}(t)\,\eta_2(x) + 2s_\epsilon\phi_\epsilon s_{\epsilon t}\eta_1(t)\,\eta_2(x) \right\}\,dx\,dt \\ &= \int_0^T\int_\Omega -k_3\left(s_\epsilon^2+\epsilon^2\right)\nabla\phi_\epsilon\cdot\nabla\eta_2(x)\,\eta_1(t)\,dx\,dt \\ &= \int_0^T\int_\Omega \left\{ -k_3 s_\epsilon^2\nabla\phi_\epsilon\cdot\nabla\eta_2(x)\,\eta_1(t) + k_3\epsilon^2\phi_\epsilon\eta_1(t)\triangle\eta_2(x) \right\}\,dx\,dt. \end{aligned} \tag{4.8}$$

By letting $\epsilon \to 0$, we can find a subsequence, $\{\epsilon_j\}_{j\in\mathbf{N}}$, such that s_{ϵ_j} converges to s weakly in $H^1\left(0,T;L^2\left(\Omega\right)\right)$, weakly-$*$ in $L^\infty\left(0,T;H^1\left(\Omega\right)\right)$ and in $L^\infty\left(\Omega_T\right)$, while ϕ_{ϵ_j} converges to a limit ϕ weakly- $*$ in $L^\infty\left(\Omega_T\right)$. Moreover, on the same subsequence, $s_{\epsilon_j}\nabla\phi_{\epsilon_j}$ converges to $s\nabla\phi$ weakly- $*$ in $L^\infty\left(0,T;L^2\left(\Omega\right)\right)$.

Now, by the uniform bounds on s_{ϵ_j}, ϕ_{ϵ_j}, and $s_{\epsilon_j}\nabla\phi_{\epsilon_j}$, every term in (4.7)–(4.8) converges to the corresponding term for s and ϕ. Note that, by Lemma 3.2, the term $s_\epsilon W'\left(s_\epsilon\right)$ is bounded uniformly in ϵ in $L^\infty\left(\Omega_T\right)$. Then, since the terms involving ϵ^2 all vanish in the limit $\epsilon \to 0$, the resulting limiting equations are the weak forms of (4.5) and (4.6). □

Remark 4.4. The equation (4.5) for the order parameter s can be obtained by multiplying the governing equation (2.9) by s. Hence, when $s \neq 0$, (2.9) and (4.5) are equivalent. On the other hand, $s \equiv 0$ is a solution of both equations.

Remark 4.5. The energy law (3.5) holds in a form of the inequality

$$E(t) + \int_0^t \int_\Omega \left\{ \left(s_t^2 + \frac{k_2}{k_3} s^2 {\phi_t}^2 \right) \right\} dx\, dt \le E(0) \tag{4.9}$$

for the weak solutions of (4.5)–(4.6) constructed in this section.

Remark 4.6. Let $s \ge 0$ (or $s \le 0$) and $0 < k_2 < k_1$. As in [11], the target space for (2.9)–(2.10) (a half of a metric cone over $\mathbf{S}^2$) can be approximated by nonpositively curved smooth manifolds. Following the arguments of [3] and taking the appropriate limit we find that the solutions of (2.9)–(2.10) are Lipschitz continuous.

5. Local existence of strong solutions. In this section we will use the semigroup approach to study the existence of strong solutions of the governing system of equations (2.9)–(2.10). Assuming that $f = (f_1, f_2)$ and $u = (s, \phi)$ we introduce the notation

$$\begin{aligned} f_1(s,\phi) &= -k_2 s |\nabla\phi|^2 - W'(s), \\ f_2(s,\phi) &= 2k_3 s^{-1} \nabla s \cdot \nabla\phi. \end{aligned} \tag{5.1}$$

In addition, we define the functional space

$$X = L^2(\Omega) \times L^2(\Omega),$$

with the norm

$$\|f\|_X^2 = \|f_1\|_{L^2(\Omega)}^2 + \|f_2\|_{L^2(\Omega)}^2.$$

We begin by rewriting the system (2.9)–(2.10) in the form

$$u_t + \mathcal{A}u = f(u, \nabla u), \qquad u = (s,\phi), \tag{5.2}$$

where f is given by (5.1) and $\mathcal{A}$ represents the linear operator

$$(\mathcal{A}u, v) = \int_\Omega (k_1 \nabla s \cdot \nabla v_1 + k_2 \nabla\phi \cdot \nabla v_2)\, dx\,.$$

Here $v = (v_1, v_2) \in H^1(\Omega) \times H^1(\Omega)$ satisfies the same boundary conditions as in (2.12). Clearly, $\mathcal{A}$ is a self-adjoint and positive definite operator. If $u \in C^2(\Omega)$ and $u = u_0$ on $\partial\Omega$, then $\mathcal{A}u = -\mathrm{div}\,(K\,\nabla u)$, where $K = \left(\begin{smallmatrix} k_1 & 0 \\ 0 & k_2 \end{smallmatrix}\right)$. Therefore,

$$\mathcal{D}(\mathcal{A}) = \left\{ u \in W^{2,2}(\Omega) \times W^{2,2}(\Omega) : u(x) = u_0,\, x \in \partial\Omega \right\}.$$

For $\alpha \geq 0$, we define

$$X^\alpha = \mathcal{D}(\mathcal{A}^\alpha), \quad \|u\|_\alpha = \|\mathcal{A}^\alpha u\|, \quad u \in X^\alpha.$$

Then we have the following lemma.

LEMMA 5.1. *Suppose that $\partial\Omega$ is a C^2-hypersurface separating $\Omega \subset \mathbf{R}^n$ from $\mathbf{R}^n \backslash \bar{\Omega}$. For $0 \leq \alpha \leq 1$, the following continuous imbedding results hold* [6]*:*

(1) $X^\alpha \subset W^{k,q},\ k - \frac{n}{q} < 2\alpha - \frac{n}{2},\ q \geq 2,$

(2) $X^\alpha \subset C^\nu,\ 0 \leq \nu < 2\alpha - \frac{n}{2},$

(3) $X^\alpha \subset L^q,\ \frac{1}{q} > \frac{1}{2} - \frac{2\alpha}{n},\ q \geq 2.$

For $\alpha \geq \frac{n}{4}$ and $\epsilon > 0$, let

$$\mathcal{U}_\epsilon = \{u \in X^\alpha : \epsilon < s < 1\}. \tag{5.3}$$

For a given $u_0 \in \mathcal{U}_\epsilon$, set

$$\mathcal{V}_\delta = \{u \in \mathcal{U}_\epsilon : \|u - u_0\|_\alpha \leq \delta\}$$

to be a δ-neighborhood of u_0 where $\delta > 0$ is fixed.

Remark 5.2. Suppose that $u \in \mathcal{V}_\delta$ for a given u_0. It follows from part (2) of Lemma 5.1 that

$$\max_{x \in \Omega \cup \partial\Omega} |s(x) - s_0(x)| \leq C(\delta),$$

where $C(\delta) \to 0$ as $\delta \to 0$. Therefore, $s(x) > s_0(x) - C(\delta) > \epsilon - C(\delta) > 0$ holds for a sufficiently small $\delta > 0$; that is, the order parameter s is uniformly bounded away from zero.

Define

$$\mu = \mu(u_0) = \epsilon - C(\delta) > 0, \qquad a = \sup\nolimits_{(s,\phi) \in \mathcal{V}_\delta} |W''(s)|;$$

then we have the following lemma.

LEMMA 5.3. *Suppose that n and α satisfy one of the following three conditions:*

$$n = 3,\ \alpha > \frac{7}{8},\ \text{ or } n = 2,\ \alpha > \frac{3}{4},\ \text{ or } n = 1,\ \alpha > \frac{5}{8}.$$

Then $F(u) = f(u, \nabla u)$ is locally Lipschitz continuous in $\mathcal{U}_\epsilon$; that is, there exists a $\delta > 0$ such that

$$\|F(u_1) - F(u_2)\|_X \leq K(\delta) \|u_1 - u_2\|_\alpha \tag{5.4}$$

for every $u_0 \in \mathcal{U}_\epsilon$ and $u_1, u_2 \in \mathcal{V}_\delta$.

Proof. For n and α satisfying the conditions of the lemma, the imbedding results of Lemma 5.1 imply

$$X^\alpha \subset W^{1,4} \cup L^\infty \quad \text{and} \quad X^\alpha \in C^\nu,$$

where

$$0 < \nu < 2\alpha - \frac{1}{2} \text{ if } n = 1,$$
$$0 < \nu < 2\alpha - 1 \text{ if } n = 2,$$
$$0 < \nu < 2\alpha - \frac{3}{2} \text{ if } n = 3.$$

Then the following estimates on (f_1, f_2) hold for every $u_1, u_2 \in \mathcal{V}_\delta$:

$$\begin{aligned}
&\left\|s_1|\nabla\phi_1|^2 - s_2|\nabla\phi_2|^2\right\|_{L^2}\\
&\quad\le \left\|(s_1 - s_2)\,|\nabla\phi_1|^2\right\|_{L^2} + \left\|s_2\left(|\nabla\phi_1|^2 - |\nabla\phi_2|^2\right)\right\|_{L^2}\\
&\quad\le \|\nabla\phi_1\|_{L^4}^2\|s_1 - s_2\|_\infty + K'\|s_2\|_\infty(\|\nabla\phi_1\|_{L^4} + \|\nabla\phi_2\|_{L^4})\|\nabla\phi_1 - \nabla\phi_2\|_{L^4}\\
&\quad\le K\|u_1 - u_2\|_\alpha
\end{aligned}$$

for some constant $K = K(\delta, u_0)$ and $K' = K'(\delta, u_0)$. Moreover,

$$\|W'(s_1) - W'(s_2)\|_{L^2} \le a\|u_1 - u_2\|_\alpha$$

and

$$\begin{aligned}
&\left\|s_1^{-1}\nabla s_1\cdot\nabla\phi_1 - s_2^{-1}\nabla s_2\cdot\nabla\phi_2\right\|_{L^2}\\
&\qquad\qquad\le \left\|s_1^{-1}\nabla s_1\cdot(\nabla\phi_1 - \nabla\phi_2)\right\|_{L^2} + \left\|\nabla\phi_2\cdot\left(s_1^{-1}\nabla s_1 - s_2^{-1}\nabla s_2\right)\right\|_{L^2}.
\end{aligned}$$

Finally, we observe that

$$\begin{aligned}
\left\|s_1^{-1}\nabla s_1\cdot(\nabla\phi_1 - \nabla\phi_2)\right\|_{L^2} &\le \mu^{-1}\|s_1\|_{W^{1,4}}\|\nabla\phi_1 - \nabla\phi_2\|_{W^{1,4}}\\
&\le \mu^{-1}K\|u_1 - u_2\|_\alpha
\end{aligned}$$

and

$$\begin{aligned}
\left\|\nabla\phi_2\cdot\left(s_1^{-1}\nabla s_1 - s_2^{-1}\nabla s_2\right)\right\|_{L^2} &\le \mu^{-2}\|\nabla\phi_2\|_{L^4}\|s_2\nabla s_1 - s_1\nabla s_2\|_{L^4}\\
&\le K\mu^{-2}\|u_2 - u_1\|_\alpha.
\end{aligned}$$

Hence we conclude that (5.4) holds. □

Remark 5.4. The estimates, similar to the ones in the proof of the previous lemma, can be used to show that

$$\|F(u)\|_X \le K(\|u\|_\infty)\|u\|_{W^{1,2k}}^k, \qquad k \ge 2, \tag{5.5}$$

where $u \in \mathcal{U}_\epsilon$ and $\|\cdot\|_\infty$ is the L^∞-norm in Ω.

THEOREM 5.5. *Let $\epsilon > 0$ and $\alpha < 1$ be as in Lemma 5.3. For any $u_0 = (s_0, \phi_0) \in \mathcal{U}_\epsilon$ there exists $T_\epsilon = T_\epsilon(u_0) > 0$ such that the boundary-value problem (2.9)–(2.12) has a unique, maximally defined solution $u(t) = (s(t), \phi(t)) \in \mathcal{D}(A)$, as long as $t \in (0, T_\epsilon)$. The solution $u(t)$ satisfies the initial conditions $s(x, 0) = s_0$ and $\phi(x, 0) = \phi_0$ for every $x \in \Omega$ and $u_i(t) \in W^{2,2}(\Omega) \cap W^{1,4}(\Omega) \cap L^\infty(\Omega)$ for $i = 1, 2$ and $t \in (0, T_\epsilon)$. Furthermore, either $T_\epsilon = \infty$ or there exists a sequence $\{t_n\} \to T_\epsilon$ such that $s(t_n) \to s_\epsilon \in \partial\mathcal{U}_\epsilon$ as $t_n \to T_\epsilon$.*

Remark 5.6. It follows from the definition of $\mathcal{U}_\epsilon$ in (5.3) that the condition $s_\epsilon \in \partial\mathcal{U}_\epsilon$ corresponds to either $\lim_{t_n\to T_\epsilon} s_\epsilon(x_0, t_n) = \epsilon$ or $\lim_{t_n\to T_\epsilon} s_\epsilon(x_0, t_n) = 1$ for some $x_0 \in \Omega$.

Remark 5.7. The preceding existence results are also valid for the initial data $s_0 \in (-\frac{1}{2}, 0)$. However, they do not apply in the case when $s_0(x) \to 0$ as $x \to \bar{x}$, where $\bar{x} \in \bar{\Omega} := \Omega \cup \partial\Omega$.

Remark 5.8. Additional regularity properties of a solution of the abstract equation (5.2) can be inferred from its definition [6]. In particular, one can show that the original equation is satisfied by this solution in a classical sense. In fact, since

$u \in D(A)$ and $u_t(\cdot, t) \in X^\alpha$ is locally Hölder continuous for $t > 0$ [6, Theorem 3.5.2], the functions u and u_t are continuous on $(0, T_\epsilon) \times \bar{\Omega}$. The inclusion $u \in D(A)$ guarantees that $\nabla u_i \in W^{1,2}(\Omega) \subset L^6(\Omega)$; hence $Au \in L^3(\Omega)$. Therefore, $u_i \in W^{2,3}(\Omega)$ and $\nabla u_i \in W^{1,3}(\Omega) \subset L^q$ for all $q > 3$. By repeating this argument, we conclude that $u_i \in W^{2,q}$ for $q > 3$ and $\nabla u(t, \cdot)$ is Hölder continuous. Then $F_i(u) \in C^\beta(\Omega)$ and $u_i(t, \cdot) \in C^{2+\beta}(\Omega)$ for some $\beta > 0$. Moreover, for $t > 0$, the function $u(x, t)$ is continuously differentiable in t and twice continuously differentiable in x; hence $u(x, t)$ is a classical solution of the governing system (2.9)–(2.10).

6. Long-time behavior and stability of solutions.

Suppose that a solution (s, ϕ) of (2.9)–(2.10) is sufficiently smooth to satisfy the energy law (cf. Theorem 3.1)

$$E'(t) + \int_\Omega \left\{ \left(s_t^2 + \frac{k_2}{k_3} s^2 {\phi_t}^2 \right) \right\} dx = 0 \tag{6.1}$$

or, in a more general case, the energy inequality (4.9)

$$E(t) + \int_0^t \int_\Omega \left\{ \left(s_t^2 + \frac{k_2}{k_3} s^2 {\phi_t}^2 \right) \right\} dx\, dt \le E(0) \tag{6.2}$$

for every $t \in (0, \infty)$. Then both terms $s_t = k_1 \triangle s - k_2 s |\nabla \phi|^2 - W'(s)$ and $s\phi_t = k_3 s^{-1} \mathrm{div}\left(s^2 \nabla \phi\right)$ are in the space $L^2\left(0, \infty; L^2(\Omega)\right)$. By using Fubini's theorem we can prove the following.

THEOREM 6.1. *Suppose that* (s, ϕ) *is a weak solution of* (2.9)–(2.12) *obtained in Theorem* 4.3. *Given any sequence* $\{t_i\}_{i \in \mathbf{N}}$ *such that* $t_i \to \infty$, *there exists a subsequence* $\{t_{i_j}\}_{j \in \mathbf{N}}$ *satisfying* $s(t_{i_j}, \cdot) \to s^*$ *in* $H^1(\Omega)$ *and* $\phi(t_{i_j}, \cdot) \overset{*}{\rightharpoonup} \phi^*$ *in* $L^\infty(\Omega)$ *weak-*$*$, *where* (s^*, ϕ^*) *is a solution of the stationary problem*

$$k_1 \triangle s - k_2 s |\nabla \phi|^2 - W'(s) = 0, \tag{6.3}$$

$$k_3 \mathrm{div}\left(s^2 \nabla \phi\right) = 0, \tag{6.4}$$

with (2.12) *as the boundary condition.*

Remark 6.2. When $0 < k_2 < k_1$ and $W \equiv 0$, one can apply the argument of R. Schoen [7], [14], [11] to show the uniqueness of the stationary minimizing solution. Then, the result of Theorem 6.1 holds not just on a subsequence but as $t \to \infty$. The same conclusion is valid when $W \neq 0$ is small in comparison with the size of the domain.

Note that there is no global (up to the boundary) uniqueness of the stationary solution when $k_2 > k_1$.

7. The exact solution in the one-dimensional case.

We now restrict Ω to lie in $\mathbf{R}$ and suppose that $\Omega := [-a, a]$. Then the system (6.3)–(6.4) can be written as

$$K^2\, s_{xx} - s\phi_x^2 - \frac{1}{\kappa^2} W'(s) = 0, \tag{7.1}$$

$$\left(s^2 \phi_x\right)_x = 0. \tag{7.2}$$

Here $K = \sqrt{k_1/k_2}$ and $\kappa = \sqrt{K_2}$. The system (7.1)–(7.2) can be decoupled by setting the angle ϕ to a constant value in Ω. Then the second equation is satisfied automatically and the first one reduces to a stationary Ginzburg–Landau equation in one dimension. We, however, will be interested in solutions of the full system with a nonconstant angular part. Therefore, we will require that $\phi(-a) < \phi(a)$. We will not impose any additional boundary conditions on s and ϕ, except for the technical restriction that $s(-a) = s(a)$. Note that, lacking uniqueness, we do not claim that we can identify the exact asymptotic limit of (2.9)–(2.10) as $t \to \infty$. In this section we will merely be interested in discussing the possible features of such a limit.

By integrating (7.2), substituting the result into (7.1), and integrating once, we obtain

$$K^2 S_x^2 + 4\alpha^2 - \frac{8}{\kappa^2} S W_1(S) = \beta S, \tag{7.3}$$

$$\phi_x = \frac{\alpha}{S}, \tag{7.4}$$

where $S := s^2$, the function $W_1(S) := W(\sqrt{S})$, and both α and β are the constants of integration. For the remainder of this section we will also assume that the potential function has a special form $W(s) = \left(1 - s^2\right)^2/8$ and, hence, $W_1(S) = (1 - S)^2/8$. This apparent relaxation of the assumptions on W from section 2 can be justified on the basis of the maximum principle result of Lemma 3.2. Then

$$(\kappa\, K\, S_x)^2 = S^3 - 2S^2 + \left(1 + \kappa^2\beta\right) S - 4\kappa^2\alpha^2. \tag{7.5}$$

The general solution of this ODE can be expressed in terms of elliptic integrals. However, for a certain combination of the constants β and α, there is a particular explicit solution of (7.5) in terms of elementary functions. To find this solution we will suppose that

$$\kappa^2\alpha^2 = \frac{1}{54}\left[1 + 9\kappa^2\beta - \left(1 - 3\kappa^2\beta\right)^{\frac{3}{2}}\right]. \tag{7.6}$$

Then (7.5) transforms into

$$(\kappa\, K\, S_x)^2 = (S - S_0(\lambda))^2\,(S - S_1(\lambda))\,, \tag{7.7}$$

where

$$S_0(\lambda) = \frac{2 + \lambda}{3}, \quad S_1(\lambda) = \frac{2(1 - \lambda)}{3},$$

and

$$\lambda = \sqrt{1 - 3\,\kappa^2\beta}. \tag{7.8}$$

Integrating (7.7) and using the restriction on the boundary values of s, we obtain

$$S(x) = S_1(\lambda) + \lambda \tanh^2\left(\frac{\sqrt{2\,\lambda}\,x}{\kappa\, K}\right). \tag{7.9}$$

We can now use (7.9) to solve (7.4):

$$\begin{aligned} \phi(x) &= \phi(-a) + \alpha \int_{-a}^{x} \frac{1}{S(p)}\, dp \\ &= \phi(-a) + \frac{\kappa\, K\, \alpha}{S_0(\lambda)\sqrt{2 S_1(\lambda)}} \left[\arctan\left\{\sqrt{\frac{\lambda}{S_1(\lambda)}} \tanh\left(\frac{\sqrt{2\,\lambda}\,x}{\kappa\, K}\right)\right\}\right. \\ &\quad \left. + \arctan\left\{\sqrt{\frac{\lambda}{S_1(\lambda)}} \tanh\left(\frac{\sqrt{2\,\lambda}\,a}{\kappa\, K}\right)\right\}\right] + \frac{\alpha}{S_0(\lambda)}(x + a) \end{aligned} \tag{7.10}$$

for every $\lambda \in (0\,,1)$. Moreover, (7.6) can be rewritten in terms of α and λ:

$$\alpha = \frac{2}{3\sqrt{3}\,\kappa}\,(2+\lambda)\,\sqrt{1-\lambda} = \frac{S_0(\lambda)\sqrt{2S_1(\lambda)}}{\kappa}. \tag{7.11}$$

Then (7.10) simplifies to

$$\begin{aligned}\phi(x) = \phi(-a) + K &\left[\arctan\left\{\sqrt{\frac{\lambda}{S_1(\lambda)}}\tanh\left(\frac{\sqrt{2\,\lambda}\,x}{\kappa\,K}\right)\right\}\right.\\ &\left.+\arctan\left\{\sqrt{\frac{\lambda}{S_1(\lambda)}}\tanh\left(\frac{\sqrt{2\,\lambda}\,a}{\kappa\,K}\right)\right\}\right] + \frac{\sqrt{2S_1(\lambda)}}{\kappa}(x+a).\end{aligned} \tag{7.12}$$

Finally, introducing a new parameter $\mu^2 = \lambda$, assuming that

$$\nu^2(\mu) = \frac{S_1(\mu^2)}{\mu^2} = \frac{2\,(1-\mu^2)}{3\,\mu^2}, \tag{7.13}$$

and using the definition of S, we have for every $\mu \in (0\,,1)$ that

$$s(x) = \mu\,\sqrt{\nu^2 + \tanh^2\left(\frac{\sqrt{2}\,\mu\,x}{\kappa\,K}\right)}, \tag{7.14}$$

$$\begin{aligned}\phi(x) = \phi(-a) + K &\left[\arctan\left\{\frac{1}{\nu}\tanh\left(\frac{\sqrt{2}\,\mu\,x}{\kappa\,K}\right)\right\}\right.\\ &\left.+\arctan\left\{\frac{1}{\nu}\tanh\left(\frac{\sqrt{2}\,\mu\,a}{\kappa\,K}\right)\right\}\right] + \frac{\sqrt{2}\,\nu\,\mu}{\kappa}\,(x+a).\end{aligned} \tag{7.15}$$

The only remaining free parameter in these equations is μ, and it can be determined by imposing the boundary conditions either on s or on ϕ. The former boundary-value problem would have a solution in the form (7.14) only if $s(-a) = s(a)$. If the boundary conditions on ϕ are given, $\phi(-a) = \phi_0$ and $\phi(a) = \phi_1$, for example, then μ can be determined from the equation

$$\phi_1 - \phi_0 = 2K\,\arctan\left\{\frac{1}{\nu(\mu)}\tanh\left(\frac{\sqrt{2}\,\mu\,a}{\kappa\,K}\right)\right\} + \frac{2\sqrt{2}\,\nu(\mu)\,\mu\,a}{\kappa}. \tag{7.16}$$

We do not claim that (7.16) would always have a solution—indeed, (7.14)–(7.15) represent a one-parameter family of solutions of the system (7.1)–(7.2) and not its general solution. However, (7.14)–(7.15) can still be used to describe some of the basic types of solutions of (7.1)–(7.2).

First, we observe that, by the definitions of λ and μ, the inequality $0 \le \mu \le 1$ always holds. Using the Maple computer algebra system [17], we will illustrate four different cases of possible singular behavior of the liquid crystal system.

(1) Suppose that $\mu = 0$. Then (7.9) and (7.12) reduce to

$$s(x) = \sqrt{\frac{2}{3}}, \tag{7.17}$$

$$\phi(x) = \phi(-a) + \frac{2}{\sqrt{3}\kappa}(x+a). \tag{7.18}$$

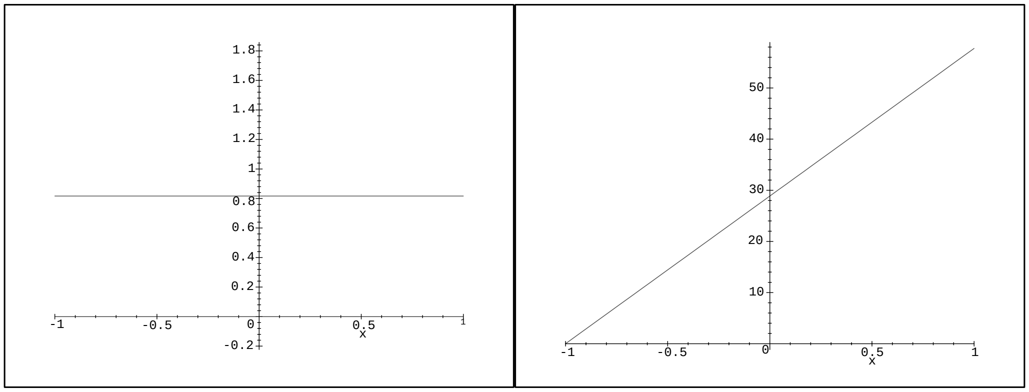


FIG. 7.1. *The graphs of s and ϕ when $K = 1$, $\mu = 0$, $\kappa = 0.04$, and $a = 1$.*

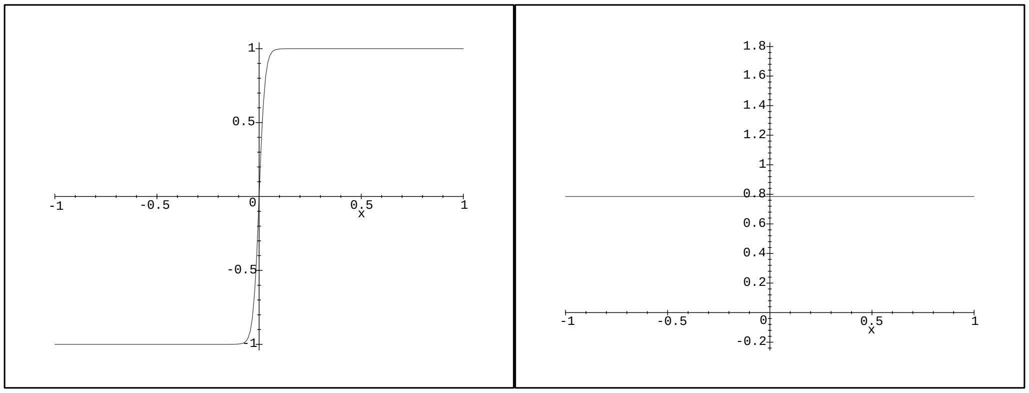


FIG. 7.2. *The graphs of s and ϕ when $K = 1$, $\mu = 1$, $\kappa = 0.04$, and $a = 1$.*

This solution (Figure 7.1) is independent of K and corresponds to a rotation of the director vector of a constant length. No structural defects are observed in this state.

(2) Now let $\mu = 1$. Then $\lambda = 1$, and by (7.11), one can see that $\alpha = 0$ holds. It follows as a consequence of (7.4) that $\phi(x)$ is constant on $(-a\,, a)$. In addition, by (7.9), we have that

$$s(x) = \tanh\left(\frac{\sqrt{2}\,x}{\kappa\,K}\right) \tag{7.19}$$

is a solution of a stationary Ginzburg–Landau equation. Typical profiles of the angle and order parameter for small values of κ are shown on Figure 7.2. Here the value of the order parameter abruptly changes at $x = 0$, while the director remains constant throughout the domain. This picture corresponds to an interface between two regions with different degrees of orientation.

(3) Here we suppose that κ is large and $\mu \to 1$. As one can observe on Figure 7.3, the corresponding physical situation can be described as a combination of the defect (wall) at $x = 0$ and the gradually changing order parameter in the bulk.

(4) We now let $\mu \to 1$ and κ be small. As μ approaches 1, the structure of the solutions changes as shown on Figures 7.4 and 7.5. The situation on Figure 7.4 can be characterized as a combination of the defect (wall) at $x = 0$ and some additional

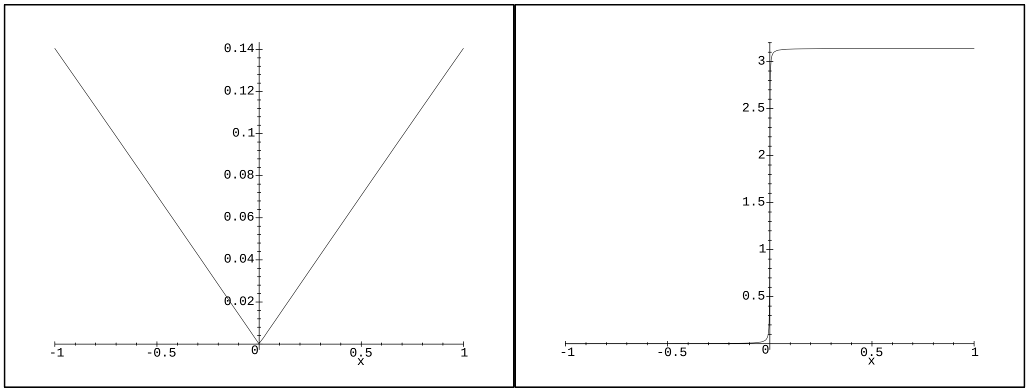


Fig. 7.3. *The graphs of s and ϕ when $K = 1$, $\mu = 0.8$, $\kappa = 10$, and $a = 1$.*

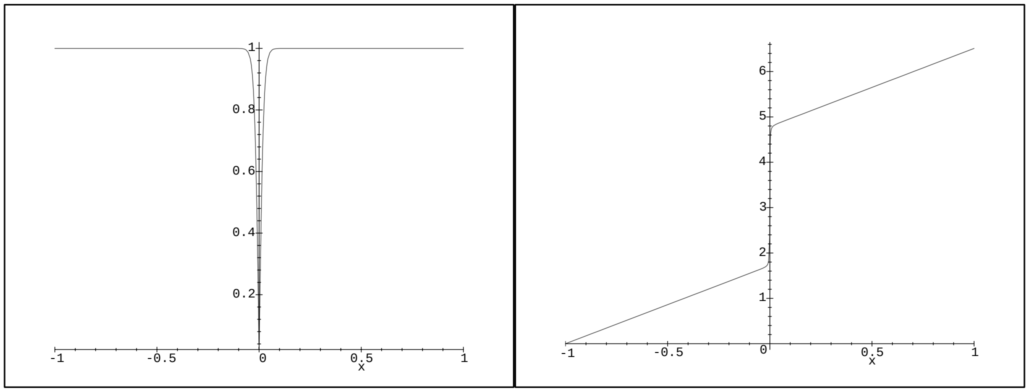


Fig. 7.4. *The graphs of s and ϕ when $K = 1$, $\mu = 99.9 \times 10^{-2}$, $\kappa = 0.04$, and $a = 1$.*

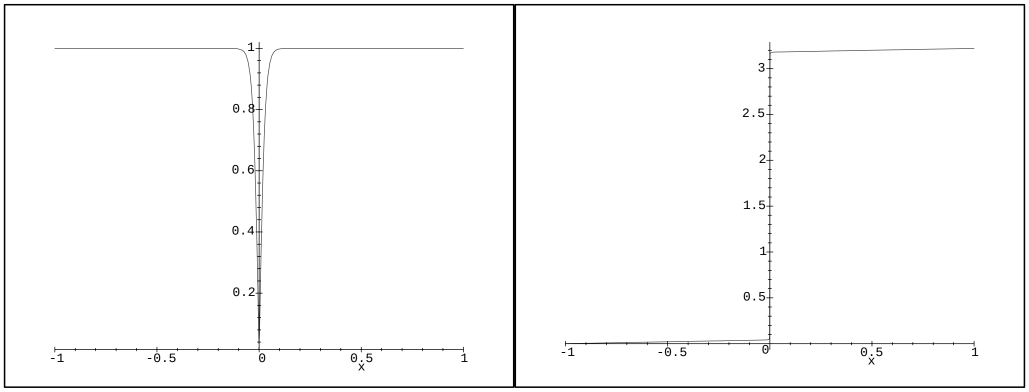


Fig. 7.5. *The graphs of s and ϕ when $K = 1$, $\mu = 99.9999 \times 10^{-2}$, $\kappa = 0.04$, and $a = 1$.*

rotation of the director in the bulk. When μ is closer to 1 (Figure 7.5), almost all of the director rotation is concentrated exclusively within the core of the defect.

We conclude, therefore, that for $\mu \in [0\,,1)$, the solution (6.14)–(6.15) is similar to a twisted nematic in which the director rotates either in the bulk or inside the defect (wall) at the center of the domain. The smallness of parameters $\frac{\kappa}{\mu}$ and $\frac{\nu}{\kappa}$ induces sharp interfaces and defects inside the domain, respectively.

8. Conclusion. We have presented the analysis of the evolution system modeling the behavior of nematic liquid crystals with variable degree of orientation. By introducing the order parameter, the system allows for the presence of structural defects in a material. This, however, creates mathematical difficulties and leads to a degenerate problem when the liquid crystal is in an isotropic state. The existence of weak solutions of the governing equations has been shown in a general case; their regularity can be established subject to certain restrictions on the order parameter. In particular, the regularity of solutions when the minima of the potential function lie on both sides of zero is still open. Using the appropriate smoothness of the director-order parameter configuration, one can then determine its asymptotic behavior and, therefore, describe the dynamics of new types of defects.

REFERENCES

[1] M. C. Calderer, *The Dynamics of Textures in Liquid Crystals*, preprint, Pennsylvania State University, University Park, PA, 1992.

[2] M. C. Calderer and C. Liu, *Liquid crystal flow: Dynamic and static configurations*, SIAM J. Appl. Math., 60 (2000), pp. 1925–1949.

[3] J. Eells, Jr., and J. H. Sampson, *Harmonic mappings of Riemannian manifolds*, Amer. J. Math., 86 (1964), pp. 109–160.

[4] J. L. Ericksen, *Liquid crystals with variable degree of orientation*, Arch. Ration. Mech. Anal., 113 (1990), pp. 97–120.

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

[6] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Springer-Verlag, Berlin, 1981.

[7] J. Jost, *Harmonic Maps between Surfaces*, Springer-Verlag, Berlin, 1984.

[8] M. Kléman, *Points, Lines, and Walls*, John Wiley and Sons, New York, 1983.

[9] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Uralceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1967.

[10] F. M. Leslie, *Theory of flow phenomena in liquid crystals*, Adv. Liquid Crystals, 4 (1979), pp. 1–81.

[11] F.-H. Lin, *On nematic liquid crystals with variable degree of orientation*, Comm. Pure Appl. Math., 44 (1991), pp. 453–468.

[12] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Springer-Verlag, New York, 1984.

[13] R. Schoen and K. Uhlenbeck, *Boundary regularity and the Dirichlet problem for harmonic maps*, J. Differential Geom., 18 (1983), pp. 253–268.

[14] M. Struwe, *Geometric evolution problems*, in Nonlinear Partial Differential Equations in Differential Geometry (Park City, UT, 1992), AMS, Providence, RI, 1996, pp. 257–339.

[15] E. G. Virga, *Defects in nematic liquid crystals with variable degree of orientation*, in Nematics (Orsay, 1990), Kluwer, Dordrecht, The Netherlands, 1991, pp. 371–390.

[16] E. G. Virga, *Variational Theories for Liquid Crystals*, Chapman and Hall, London, 1994.

[17] Waterloo Maple Inc., *Maple* V, release 4, Waterloo, ON, Canada, 1996.

PERMANENCE OF WEAKLY COUPLED VECTOR FIELDS*

SEBASTIAN J. SCHREIBER†

Abstract. Let $F^1, \dots, F^k$ be k dissipative vector fields on finite dimensional Euclidean spaces that preserve the skeleton of the positive orthant. Permanence of all sufficiently weak couplings of these vector fields corresponds to robust permanence of the uncoupled vector field $F^1 \times \cdots \times F^k$. A sufficient condition for robust permanence of $F^1 \times \cdots \times F^k$ involving unsaturated Morse decompositions is provided. In the case of coupled food chain vector fields and coupled two-dimensional vector fields, this sufficient condition is shown to be necessary. As an illustration, these results are applied to weakly coupled logistic-Holling predator-prey systems.

Key words. permanence, weak coupling, population dynamics

AMS subject classifications. 92D15, 60J10, 60F99

PII. S0036141001387434

1. Introduction.

Ecological communities are bound together by a web of complex relationships. Data on interaction strengths between species in natural food webs indicate that these interaction strengths are characterized by many weak interactions and a few strong interactions [2, 19, 21, 30]. Consequently, food webs are often viewed as modular [22] in which modules of strongly interacting species are weakly interconnected. Because of this modularity, theoretical ecologists often develop ordinary differential equation models that include strong interactions and ignore weak interactions [11, 24, 25]. One justification for this approach is the belief that weak interactions play a small role on the models dynamics and therefore can be ignored. More recently, however, theoretical ecologists have considered the effects of weak interactions and have found that weak interactions can play important stabilizing and "noise" dampening roles [2, 19, 21] and can magnify spatiotemporal variation in community structure [2]. Furthermore, weak couplings of predator-prey systems with periodic orbits exhibit phase locking and entrainment [28]. To better understand their global dynamics, we investigate permanence of weakly coupled systems.

Population dynamics are frequently modeled by vector fields on the positive orthant of Euclidean space that leave the boundary of this orthant invariant. Stated loosely, such a vector field is permanent provided that the boundary of the positive orthant is repelling [3, 4, 9, 13, 14, 27]. Ecologically this is interpreted as the long-term coexistence of the interacting populations. Reviews of the mathematical progress on studying permanence and its applications can be found in [10, 15, 29]. If we have k-vector fields $F^1, \dots, F^k$ representing k uncoupled systems of interacting populations, then weak couplings of these vector fields correspond to vector fields that are close to the product vector field $F^1 \times \cdots \times F^k$. The goal of this article is to find conditions that ensure all vector fields sufficiently close to the uncoupled vector field $F^1 \times \cdots \times F^k$ are permanent. In other words, determine under what conditions $F^1 \times \cdots \times F^k$ is robustly permanent. After defining these concepts more precisely in section 2, we conjecture that $F^1 \times \cdots \times F^k$ is robustly permanent if and only if for each $1 \le i \le k$ the vector field F^i is robustly permanent. If this conjecture is indeed true, then it reduces a

*Received by the editors April 5, 2001; accepted for publication (in revised form) August 15, 2001; published electronically January 16, 2002. This work was supported by NSF grant DMS-0077986.

http://www.siam.org/journals/sima/33-5/38743.html

†Department of Mathematics, Western Washington University, Bellingham, WA 98225 (sschreib@cc.wwu.edu).

potentially high-dimensional problem (i.e., determining whether $F^1 \times \cdots \times F^k$ has a robustly repelling boundary) to k lower-dimensional problems (i.e., determining for each $1 \leq i \leq k$ whether F^i has a robustly repelling boundary). To prove such a conjecture requires an appropriate characterization of robust permanence and extending it to product vector fields. Results in this direction include the recent work of the author [26] in which it is shown that a vector field is robustly permanent provided that it admits an unsaturated Morse decomposition of its boundary dynamics, and a generalization of this criterion by Hirsch, Smith, and Zhao [8] to semiflows. Recall that a Morse decomposition of an invariant set K is a collection of isolated invariant subsets of K such that collapsing these sets to distinct points results in a gradient-like quotient flow on K. On the other hand, an unsaturated equilibrium in a population model is in effect one which can be invaded by some population entering the system at low density. Thus an unsaturated Morse decomposition is one in which the invariant subsets in the decomposition, which may be more complicated than just equilibria, have the analogous property of invasibility by at least one population not already present. In section 3, we discuss these results and prove that if $F^1, \ldots, F^k$ are vector fields that admit an unsaturated Morse decomposition, then $F^1 \times \cdots \times F^k$ admits an unsaturated Morse decomposition and, consequently, is robustly permanent. In section 4, we use the main result of section 3 to prove our conjecture for couplings of two-dimensional vector fields and for couplings of food chain vector fields. In section 5, we prove a technical proposition to make our results more applicable and illustrate our approach with coupled predator-prey systems of the logistic-Holling type.

2. Preliminaries and a conjecture.

DEFINITION 2.1. *Let $\mathcal{P}^r(n)$ be the space of C^r vector fields $F = (F_1, \ldots, F_n) : \mathbf{R}^n_+ \to \mathbf{R}^n$ that satisfy $F_i(x) = 0$ whenever $x_i = 0$.*

The extra condition on F corresponds to the fact that in the absence of population i, the growth rate of population i is zero. We view $\mathcal{P}^r(n)$ as the space of all possible models of n-interacting populations and endow $\mathcal{P}^r(n)$ with the C^r Whitney topology [7, Chapter 2].

DEFINITION 2.2. *Let $F = (F_1, \ldots, F_n) \in \mathcal{P}^r(n)$ with $r \geq 1$. The* per capita growth functions *$f = (f_1, \ldots, f_n) : \mathbf{R}^n_+ \to \mathbf{R}^n$ associated with F are the continuous functions defined by*

$$f_i(x) = \begin{cases} \frac{F_i(x)}{x_i} & \text{if } x_i \neq 0, \\ \frac{\partial F_i}{\partial x_i}(x) & \text{if } x_i = 0 \end{cases} \tag{2.1}$$

for any $x = (x_1, \ldots, x_n) \in \mathbf{R}^n_+$.

We recall several definitions from dynamical systems theory. Assume $F : \mathbf{R}^n_+ \to \mathbf{R}^n$ is C^1 and that $\dot{x} = F(x)$ generates a global flow $\phi : \mathbf{R} \times \mathbf{R}^n_+ \to \mathbf{R}^n_+$. Let $\phi_t x = \phi(t, x)$. Given sets $I \subseteq \mathbf{R}$ and $K \subseteq \mathbf{R}^n_+$, let $\phi_I K = \{\phi_t x : t \in I, x \in K\}$. A set $K \subseteq \mathbf{R}^n_+$ is called *invariant* if $\phi_t K = K$ for all $t \in \mathbf{R}$. The *omega limit set* of a set $K \subseteq \mathbf{R}^n_+$ equals $\omega(F, K) = \cap_{t \geq 0} \overline{\phi_{[t,\infty)} K}$. The *alpha limit set* of a set $K \subseteq \mathbf{R}^n_+$ equals $\alpha(F, K) = \cap_{t \leq 0} \overline{\phi_{(-\infty,t]} K}$. $A \subset \mathbf{R}^n_+$ is called an *attractor* for ϕ provided there exists an open neighborhood $U \subseteq \mathbf{R}^n_+$ of A such that $\omega(F, U) = A$. The *basin of attraction* of A is the set of points $x \in \mathbf{R}^n_+$ such that $\omega(F, x) \subseteq A$. The flow ϕ is *dissipative* if there exists a compact attractor $A \subset \mathbf{R}^n_+$ for ϕ whose basin of attraction is $\mathbf{R}^n_+$.

DEFINITION 2.3. *$F \in \mathcal{P}^r(n)$ is* permanent *provided that $\dot{x} = F(x)$ generates a dissipative flow ϕ and there exists a compact attractor $A \subset \text{int}\mathbf{R}^n_+$ for ϕ whose basin of attraction is $\text{int}\mathbf{R}^n_+$.*

Permanence was originally introduced in [27] and for dissipative vector fields is equivalent to uniform persistence [4]. To talk about permanence of weakly coupled vector fields, the following definition is useful.

DEFINITION 2.4. *$F \in \mathcal{P}^r(n)$ is C^r* robustly permanent *if there exists a neighborhood $\mathcal{N} \subseteq \mathcal{P}^r(n)$ of F such that every vector field $G \in \mathcal{N}$ is permanent.*

Now suppose that $F^1 \in \mathcal{P}^r(n_1), \ldots, F^k \in \mathcal{P}^r(n_k)$ are vector fields. Define the product vector field $F^1 \times \cdots \times F^k \in \mathcal{P}^r(n_1 + \cdots + n_k)$ by

$$F^1 \times \cdots \times F^k(x^1, \ldots, x^k) = (F^1(x^1), \ldots, F^k(x^k)),$$

where $x^1 \in \mathbf{R}_+^{n_1}, \ldots, x^k \in \mathbf{R}_+^{n_k}$. If $F^1 \times \cdots \times F^k$ is C^r robustly permanent, then all sufficiently weak C^r couplings of $F^1, \ldots, F^k$ are permanent.

We make the following conjecture.

CONJECTURE 1. *Let $F^1 \in \mathcal{P}^r(n_1), \ldots, F^k \in \mathcal{P}^r(n_k)$ with $r \geq 1$ be vector fields that generate dissipative flows. If F^i is C^r robustly permanent for each $1 \leq i \leq k$, then $F^1 \times \cdots \times F^k$ is C^r robustly permanent.*

The converse of this conjecture—namely, robust permanence of $F^1 \times \cdots \times F^k$ implies robust permanence of F^i for each $1 \leq i \leq k$—follows immediately from the definitions. The utility of this conjecture, provided that it is true, is that it reduces checking robust permanence of a $n_1 + \cdots + n_k$-dimensional vector field to checking robust permanence of n_i-dimensional vector fields for $1 \leq i \leq k$.

3. Unsaturated Morse decompositions. In previous work [26], the author developed a sufficient condition for robust permanence. This condition involves the notion of unsaturated Morse decompositions that we discuss now. Let $F \in \mathcal{P}^r(n)$ with $r \geq 1$ generate the dissipative flow ϕ and let $f = (f_1, \ldots, f_n)$ denote the per capita growth rate functions associated with F. Given a compact invariant set K, let $\mathcal{M}_{\text{inv}}(F, K)$ denote the set of ϕ-invariant Borel probability measures with support in K. A compact invariant set K for ϕ is *unsaturated* if

$$\min_{\mu \in \mathcal{M}_{\text{inv}}(F,K)} \max_{1 \leq i \leq n} \int_K f_i \, d\mu > 0.$$

Recall that a compact invariant set K is *isolated* if there exists a neighborhood V of K such that K is the maximal compact invariant set in V. A collection of sets $\{M_1, \ldots, M_k\}$ is a *Morse decomposition* for a compact invariant set K if $M_1, \ldots, M_k$ are pairwise disjoint, compact isolated invariant sets for $\phi|K$ with the property that for each $x \in K$ there are integers $l = l(x) \leq m = m(x)$ such that $\alpha(F, x) \subseteq M_m$ and $\omega(F, x) \subseteq M_l$ and if $l = m$, then $x \in M_l = M_m$. Let K be a compact invariant set. We say that $\{M_1, \ldots, M_k\}$ is an *unsaturated Morse decomposition for K* if $\{M_1, \ldots, M_k\}$ is a Morse decomposition for K and each M_j is unsaturated. The following sufficient condition for C^r robust permanence was proven by the author.

THEOREM 3.1 (Schreiber [26]). *Let $F \in \mathcal{P}^r(n)$ with $r \geq 1$ be such that $\dot{x} = F(x)$ generates a dissipative flow ϕ. Let $\Lambda \subset \partial \mathbf{R}_+^n$ be the maximal compact invariant set for $\phi|\partial \mathbf{R}_+^n$. If Λ admits an unsaturated Morse decomposition, then F is C^r robustly permanent.*

As a step toward our conjecture, we prove a direct product of flows that admit unsaturated Morse decompositions admits an unsaturated Morse decomposition.

THEOREM 3.2. *If $F^1 \in \mathcal{P}^r(n_1), \ldots, F^k \in \mathcal{P}^r(n_k)$ with $r \geq 1$ are vector fields such that for each $1 \leq i \leq k$*

1. *F^i generates a dissipative flow,*

2. *F^i admits an unsaturated Morse decomposition of the maximal compact invariant set of $\partial\mathbf{R}_+^{n_i}$,*

then $F^1 \times \cdots \times F^k$ is C^r robustly permanent.

The proof of this theorem follows from induction, the following proposition, and the fact that dissipative vector fields are open in the C^1 Whitney topology.

PROPOSITION 3.3. *Let $F \in \mathcal{P}^r(n)$ and $G \in \mathcal{P}^r(m)$ with $r \geq 1$ generate dissipative flows ϕ_t and ψ_t, respectively. Let $\Lambda_1 \subset \mathbf{R}_+^n$ and $\Lambda_2 \subset \mathbf{R}_+^m$ be the maximal compact invariant sets for ϕ and ψ, respectively. If $\Lambda_1 \cap \partial\mathbf{R}_+^n$ admits an unsaturated Morse decomposition for ϕ and $\Lambda_2 \cap \partial\mathbf{R}_+^m$ admits an unsaturated Morse decomposition for ψ, then $\Lambda_1 \times \Lambda_2 \cap \partial\mathbf{R}_+^{n+m}$ admits an unsaturated Morse decomposition for $\phi \times \psi$.*

Proof. Let $\{N_1, \ldots, N_k\}$ and $\{M_1, \ldots, M_l\}$ be unsaturated Morse decompositions for $\phi|\Lambda_1 \cap \partial\mathbf{R}_+^n$ and $\psi|\Lambda_2 \cap \partial\mathbf{R}_+^m$, respectively. Theorem 3.1 implies that ϕ and ψ are permanent. Hence there exist compact attractors $N_0 \subset \mathrm{int}\mathbf{R}_+^n$ and $M_0 \subset \mathrm{int}\mathbf{R}_+^m$ for ϕ and ψ, respectively, whose basins of attraction are $\mathrm{int}\mathbf{R}_+^n$ and $\mathrm{int}\mathbf{R}_+^m$, respectively. Hence $\{N_0, \ldots, N_k\}$ and $\{M_0, \ldots, M_l\}$ define Morse decompositions for $\phi|\Lambda_1$ and $\psi|\Lambda_2$, respectively.

Define the collection of sets $\{V_i\}_{i=0}^{l(k+1)+k}$ by

$$V_{j(k+1)+i} = N_i \times M_j, \qquad i = 0, \ldots, k, \quad j = 0, \ldots, l.$$

We claim that this collection of sets is a Morse decomposition for $\phi \times \psi$ restricted to $\Lambda_1 \times \Lambda_2$. Given any $1 \leq i \leq k$ and $1 \leq j \leq k$, N_i and M_j have isolating neighborhoods for $\phi|\Lambda_1$ and $\psi|\Lambda_2$. The product of these isolating neighborhoods is an isolating neighborhood of $N_i \times M_j$ for $\phi \times \psi|\Lambda_1 \times \Lambda_2$. Hence, $\{V_i\}_{i=0}^{l(k+1)+k}$ is a collection of isolated invariant sets for $\phi \times \psi|\Lambda_1 \times \Lambda_2$. Given $z = (x, y) \in \Lambda_1 \times \Lambda_2$, there exist $0 \leq i \leq k$ and $0 \leq j \leq l$ such that $\alpha(F, x) \subset N_i$ and $\alpha(G, y) \subset M_j$. Hence, $\alpha(F \times G, z) \subset V_{j(k+1)+i}$. Since $\{N_1, \ldots, N_k\}$ and $\{M_1, \ldots, M_l\}$ are Morse decompositions for F and G, respectively, there exist $i' \leq i$ and $j' \leq j$ such that $\omega(F, x) \subset M_{i'}$ and $\omega(G, y) \subset N_{j'}$. Furthermore, $i' = i$ only if $x \in N_i$ and $j' = j$ only if $y \in M_j$. Consequently, $\omega(F \times G, z) \subset N_{i'} \times M_{j'} = V_{j'(k+1)+i'}$ with $j'(k+1) + i' \leq j(k+1) + i$. Furthermore, $j'(k+1) + i' = j(k+1) + i$ only if $z \in V_{j(k+1)+i}$. Hence, $\{V_i\}_{i=0}^{l(k+1)+k}$ is a Morse decomposition for $\phi \times \psi|\Lambda_1 \times \Lambda_2$.

Since $V_0 \subset \mathrm{int}\mathbf{R}_+^{n+m}$ is an attractor for $\phi \times \psi$ with basin of attraction $\mathrm{int}\mathbf{R}_+^{n+m}$, $\{V_i\}_{i=1}^{l(k+1)+k}$ defines a Morse decomposition for $\phi \times \psi$ restricted to $\Lambda = (\Lambda_1 \times \Lambda_2) \cap \partial\mathbf{R}_+^{n+m}$. We claim that this Morse decomposition is unsaturated. Let $(f_1, \ldots, f_n)$ and $(g_1, \ldots, g_l)$ be the per capita growth rate functions of F and G, respectively. The per capita growth rate functions for $H = F \times G$ are given by

$$(h_1(x,y), \ldots, h_n(x,y), h_{n+1}(x,y), \ldots, h_{n+m}(x,y)) = (f_1(x), \ldots, f_n(x), g_1(y), \ldots, g_m(y)).$$

Let $V_{j(k+1)+i} = N_i \times M_j$ with $0 \leq i \leq k$ and $0 \leq j \leq l$ be a component of this Morse decomposition. Due to the fact that $j(k+1)+i \geq 1$, we have that either $i \neq 0$ or $j \neq 0$. Assume that $i \neq 0$ as the case $j \neq 0$ can be treated similarly. Since $\mathcal{M}_{\mathrm{inv}}(H, N_i \times M_j)$ is compact in the weak* topology, to show that $V_{j(k+1)+i}$ is unsaturated reduces to verifying that

$$\max_{1 \leq l \leq n+m} \int_{N_i \times M_j} h_l \, d\mu > 0$$

for every $\mu \in \mathcal{M}_{\mathrm{inv}}(H, N_i \times M_j)$. Let $\mu \in \mathcal{M}_{\mathrm{inv}}(H, N_i \times M_j)$. Define $\pi : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$ by $\pi(x, y) = x$. Let $\pi_*\mu$ be the Borel probability measure on N_i defined by

$\pi_*\mu(A) = \mu(\pi^{-1}(A))$ for all Borel sets $A \subset N_i$. $\pi_*\mu$ is invariant for ϕ as for any continuous function $c : N_i \to \mathbf{R}$ and $t \in \mathbf{R}$,

$$\int_{N_i} c \circ \phi_t \, d\pi_*\mu = \int_{N_i \times M_j} c \circ \phi_t \circ \pi \, d\mu = \int_{N_i \times M_j} c \circ \pi \circ \phi_t \times \psi_t \, d\mu$$
$$= \int_{N_i \times M_j} c \circ \pi \, d\mu = \int_{N_i} c \, d\pi_*\mu$$

where the third equality is given by the invariance of μ. Since M_i is unsaturated for ϕ,

$$\max_{1 \le l \le n} \int_{N_i} f_l \, d\pi_*\mu > 0. \tag{3.1}$$

Since $\int_{N_i \times M_j} h_l \, d\mu = \int_{N_i} f_l \, d\pi_*\mu$ for $1 \le l \le n$, (3.1) implies that

$$\max_{1 \le l \le n+m} \int_{N_i \times M_j} h_l \, d\mu > 0.$$

Hence $N_i \times M_j$ is unsaturated for $\phi \times \psi$. □

4. Two corollaries.

In the next two subsections, we derive two corollaries of Theorem 3.2.

4.1. Weakly coupled two-dimensional systems.

The basic building blocks of ecological theory are two-species interactions that include competition, mutualism, and predator-prey interactions [1]. The following corollary of Theorem 3.2 implies that our conjecture is true for couplings of two-dimensional vector fields.

COROLLARY 4.1. *Let $F^1 \in \mathcal{P}^r(2), \dots, F^k \in \mathcal{P}^r(2)$ with $r \ge 1$ be vector fields that generate dissipative flows. If (f_1^i, f_2^i) denote the per capita growth rates of F^i for $1 \le i \le k$, then $F^1 \times \cdots \times F^k$ is robustly permanent if and only if for each $1 \le i \le k$*

1. *$f_1^i(0) > 0$ or $f_2^i(0) > 0$,*
2. *$f_2^i(x) > 0$ for all $x = (x_1, 0) \in \mathbf{R}_+^2 \backslash \{0\}$ such that $f_1^i(x) = 0$, and*
3. *$f_1^i(x) > 0$ for all $x = (0, x_2) \in \mathbf{R}_+^2 \backslash \{0\}$ such that $f_2^i(x) = 0$.*

Remark. Since two-dimensional vector fields can exhibit periodic orbits as well as equilibria, the boundary dynamics of the direct product of several two-dimensional vector fields may exhibit periodic and quasi-periodic orbits. Despite these complications, the corollary implies that robust permanence of uncoupled two-dimensional vector fields is determined by easily verified conditions at equilibria.

Proof. First, suppose that each F^i for $1 \le i \le k$ satisfies the three conditions of the corollary. For each $1 \le i \le k$, we will show that F^i admits an unsaturated Morse decomposition. Let Λ^i be the global attractor for F^i. Condition 1 implies that either $f_1^i(0) > 0$ or $f_2^i(0) > 0$. Without loss of generality, assume that $f_1^i(0) > 0$. Define $M_1 = \Lambda^i \cap \{(x_1, 0) : x_1 > 0\}$ and $M_2 = \Lambda^i \cap \{(0, x_2) : x_2 \ge 0\}$. Since $f_1^i(0) > 0$, M_1 and M_2 are disjoint isolated invariant sets for $\partial\mathbf{R}_+^2$. Invariance of $\{(0, x_2) : x_2 \ge 0\}$ implies that every point $x = (0, x_2) \in \Lambda^i$ satisfies $\alpha(F^i, x) \cup \omega(F^i, x) \subset M_2$. Invariance of $\{(x_1, 0) : x_1 \ge 0\}$ and the fact that $f_1^i(0) > 0$ implies that every point $x = (x_1, 0) \in \mathbf{R}_+^2$ with $x_1 > 0$ satisfies $\omega(F^i, x) \subset M_1$. Hence, $\{M_1, M_2\}$ is a Morse decomposition for F^i restricted to $\Lambda^i \cap \partial\mathbf{R}_+^2$. Since all invariant measures for F^i restricted to $\partial\mathbf{R}_+^2$ are convex combinations of Dirac measures based at equilibria, conditions 1–3 imply that $\{M_1, M_2\}$ is an unsaturated Morse decomposition for F^i restricted to $\Lambda^i \cap \partial\mathbf{R}_+^2$. Theorem 3.2 implies that $F_1 \times \cdots \times F_k$ is C^r robustly permanent.

On the other hand, suppose that one of the conditions is not met for one of the vector fields F^i. We will show that for every C^r neighborhood $\mathcal{N} \subset \mathcal{P}^r(2)$ of F^i contains a vector field that is not permanent. It follows that $F^1 \times \cdots \times F^k$ is not C^r robustly permanent. If condition 1 is not met for F^i, then there is a $G \in \mathcal{N}$ such that the origin is linearly stable for G. If condition 2 is not met for F^i, then there is an equilibrium $x^* = (x_1^*, 0)$ such that $f_2^i(x^*) \leq 0$. An appropriate perturbation of F^i yields a $G \in \mathcal{P}^r(2)$ such that x^* is an equilibrium for G and $g_2(x^*) < 0$, where (g_1, g_2) are the per capita growth rate functions for G. The stable manifold theorem implies there is a point $y \in \text{int}\mathbf{R}_+^2$ such that $\omega(y) = x^*$. Hence G is not permanent. Similarly, if condition 3 is not met for F^i, then we can find $G \in \mathcal{N}$ such that G is not permanent. □

4.2. Weakly coupled food chains. Using Theorem 3.2, we can prove our conjecture for food chain vector fields that represent a collection of populations where the ith population consumes the $(i-1)$st population and is consumed by the $(i+1)$st population [1]. Food chain models represent a fundamental ecological unit whose dynamics have been studied extensively [5, 6, 16, 17, 18, 20, 23].

Definition 4.2. *Let $F \in \mathcal{P}^r(n)$ with $r \geq 1$ and $f = (f_1, \ldots, f_n)$ denote the per capita growth rate functions. F is a* food chain vector field *provided that F generates a dissipative flow and for all $2 \leq i \leq n$, $f_i(x) < 0$ whenever $x_{i-1} = 0$.*

The definition of a food chain vector field asserts that in the absence of the $(i-1)$st population for $i \geq 2$, the ith population has a negative per capita growth rate and is doomed to extinction. Population 1 plays a special role under this assumption as $f_1(0)$ is permitted to be positive. Population 1 in food chain models typically represents an auto-trophic population (e.g., a population of plants) whose resources are not explicitly modeled.

Given a vector field F, recall a point $x \in \mathbf{R}_+^n$ is *recurrent* if $x \in \omega(F, x)$. The *Birkhoff center* for a compact invariant set K, denoted $BC(F, K)$, is the closure of the recurrent points of K. The following characterization of C^r robust permanence for food chain models was proven by the author.

Theorem 4.3 (Schreiber [26]). *Let $F \in \mathcal{P}^r(n)$ with $r \geq 1$ be a food chain vector field with $r \geq 1$ that generates a dissipative flow ϕ. Let Λ be the maximal compact invariant set for $\phi|\partial\mathbf{R}_+^n$. Then the following are equivalent:*

1. *F is C^r robustly permanent.*
2. *There exist compact sets $A_0 = \{0\} = \mathbf{R}_+^0, A_1 \subset \text{int}\mathbf{R}_+^1, \ldots, A_{n-1} \subset \text{int}\mathbf{R}_+^{n-1}$ and $t > 0$ such that for each $m \in \{0, 1, \ldots, n-1\}$, A_m is an attractor for $\phi|\mathbf{R}_+^m$ with basin of attraction $\text{int}\mathbf{R}_+^m$ and*

$$\min_{x \in BC(F, A_m)} \int_0^t f_{m+1}(\phi_s x)\,ds > 0, \tag{4.1}$$

where $BC(F, A_m)$ is the Birkhoff center of $\phi|A_m$. In particular, $\{A_0, \ldots, A_{n-1}\}$ defines an unsaturated Morse decomposition for $\phi|\Lambda$.

With this characterization in hand, we immediately get the following corollary.

Corollary 4.4. *If $F^1 \in \mathcal{P}^r(n_1), \ldots, F^k \in \mathcal{P}^r(n_k)$ are food chain vector fields with $r \geq 1$, then $F^1 \times \cdots \times F^k$ is C^r robustly permanent if and only if F^i is robustly permanent for each $1 \leq i \leq k$.*

5. Using the results. Although the C^r Whitney topology on $\mathcal{P}^r(n)$ is natural from a theoretical perspective, the perturbations considered in most models are not

C^r close in this strong sense. Fortunately, a means to bypass this issue is contained in the following proposition.

PROPOSITION 5.1. *Let $F \in \mathcal{P}^r(n)$ with $r \geq 1$ be C^r robustly permanent. Let $\Lambda \subset \mathbf{R}^n_+$ be the maximal compact invariant set for F. Let V_1 be a compact neighborhood of Λ and V_2 a compact neighborhood of V_1. Then there exists $\epsilon > 0$ such that if $G \in \mathcal{P}^r(n)$ satisfies*

1. *G generates a dissipative flow,*
2. *the maximal compact invariant set for G is contained in V_1, and*
3. *$\|G(x) - F(x)\| + \|DG(x) - DF(x)\| + \cdots + \|D^rG(x) - D^rF(x)\| \leq \epsilon$ for all $x \in V_2$,*

then G is permanent.

Proof. Let F be C^r robustly permanent and have maximal compact invariant set Λ. Since F is C^r robustly permanent, there exists a C^r neighborhood $\mathcal{N} \subset \mathcal{P}^r(n)$ of F such that every $G \in \mathcal{N}$ is permanent. Without loss of generality, we may assume that this neighborhood is given by

$$\{G \in \mathcal{P}^r(n) : \|G(x) - F(x)\|_r < c(x) \text{ for all } x \in \mathbf{R}^n_+\},$$

where $c : \mathbf{R}^n_+ \to (0,1]$ is a continuous function and

$$\|G(x) - F(x)\|_r = \|G(x) - F(x)\| + \|DG(x) - DF(x)\| + \cdots + \|D^rG(x) - D^rF(x)\|.$$

Let $\rho : \mathbf{R}^n_+ \to [0,1]$ be a smooth function such that $\rho(x) = 1$ for all $x \in V_1$ and $\rho(x) = 0$ for all $x \in \mathbf{R}^n_+ \setminus V_2$. Define

$$\epsilon = \min_{x \in V_2} \frac{c(x)}{(r+2)!(1 + \|\rho(x)\|_r)}.$$

Let $G \in \mathcal{P}^r(n)$ be a vector field that satisfies conditions in the statement of the proposition. Define $\tilde{G}(x) = F(x) + \rho(x)(G(x) - F(x))$. We claim that $\tilde{G}$ lies in $\mathcal{N}$. To this end, we need to make estimates for $\|D^i\tilde{G}(x) - D^iF(x)\| = \|D^i(\rho(x)(G(x) - F(x))\|$ for all $x \in \mathbf{R}^n_+$ and $0 \leq i \leq r$. The product rule implies that for $x \in V_2$ and $0 \leq i \leq r$,

$$\begin{aligned}\|D^i\left(\rho(x)(G(x) - F(x))\right)\| &\leq i! \sum_{j=0}^{i} \|D^{j-i}\rho(x) D^j(G(x) - F(x))\| \\ &\leq (i+1)! \|\rho(x)\|_r\, \epsilon.\end{aligned}$$

Consequently, for $x \in V_2$,

$$\|\tilde{G}(x) - F(x)\|_r \leq (r+2)! \|\rho(x)\|_r\, \epsilon \leq c(x).$$

On the other hand, for $x \in \mathbf{R}^n_+ \setminus V_2$, $\|F(x) - \tilde{G}(x)\|_r = 0$. Hence $\tilde{G}$ lies in $\mathcal{N}$ and is permanent. Since $\tilde{G} = G$ on the maximal compact invariant set for G, it follows that G is permanent. □

Now we illustrate how these results can be used for coupled predator-prey systems similar to those considered by Vandermeer [28]. Vandermeer used MacArthur predator-prey equations, but we choose not to use these equations as they exhibit some pathological properties (i.e., the equations are not continuous on the boundary of phase space). Instead we assume that the prey exhibits logistic dynamics in the absence of the predator, and the predator has a Holling type II function response [1]. Coupling these equations only through the predators, we get

$$\begin{aligned}
&\frac{dx_i}{dt} = r_i x_i\left(1-\frac{x_i}{K_i}\right) - \frac{a_i x_i y_i}{1+b_i x_i+\epsilon c_i x_j} - \frac{\epsilon d_j y_j x_i}{1+b_j x_j+\epsilon c_j x_i},\\
&\frac{dy_i}{dt} = \frac{e_i x_i y_i}{1+b_i x_i+\epsilon c_i x_j} + \frac{\epsilon f_i y_i x_j}{1+b_i x_i+\epsilon c_i x_j} - m_i y_i, \qquad (5.1)\\
&i,j\in\{1,2\},\quad i\neq j,
\end{aligned}$$

where x_i and y_i are the densities of prey species i and predator species i, respectively. The parameters have the following interpretation: r_i is the intrinsic rate of growth of prey species i; K_i is the carrying capacity of prey species i; a_i and ϵd_i correspond to the rates at which the predators encounter prey species; $\frac{b_i}{a_i}$ and $\frac{c_i}{d_i}$ correspond to prey handling times for predator species i; $\frac{e_i}{a_i}$, $\frac{f_i}{d_i}$ correspond to the conversion rate of prey numbers to predator numbers; and m_i is the per capita mortality rate of predator species i. $\epsilon = 0$ corresponds to the uncoupled system. The fact that this coupling is somewhat complex follows from the fact that the predators are handling both prey species when $\epsilon > 0$.

THEOREM 5.2. *Let $r_i, K_i, a_i, b_i, c_i, d_i, e_i, f_i$, and m_i be positive reals. There exists an $\tilde{\epsilon} > 0$ such that* (5.1) *is permanent for all* $0 \le \epsilon < \tilde{\epsilon}$ *if and only if* $K_i > \frac{m_i}{e_i - b_i m_i} > 0$ *for* $i = 1, 2$.

Proof. Consider the uncoupled system (i.e., $\epsilon = 0$ in (5.1)) that is given by

$$\begin{aligned}
&\frac{dx_i}{dt} = r_i x_i\left(1-\frac{x_i}{K_i}\right) - \frac{a_i x_i y_i}{1+b_i x_i},\\
&\frac{dy_i}{dt} = \frac{e_i x_i y_i}{1+b_i x_i} - m_i y_i, \qquad i = 1,2. \qquad (5.2)
\end{aligned}$$

The only boundary equilibria for the predator-prey subsystem $x_i - y_i$ of (5.2) are given by the origin $(x_i, y_i) = (0, 0)$ and $(x_i, y_i) = (K_i, 0)$. These equilibria satisfy the conditions of Corollary 4.1 if and only if $K_i > \frac{m_i}{e_i - b_i m_i} > 0$. Alternatively, if $K_i \le \frac{m_i}{e_i - b_i m_i}$ or $\frac{m_i}{e_i - b_i m_i} \le 0$ for some $i \in \{1, 2\}$, it can be shown [12, Lemma 3.2] that predator i is driven to extinction in the uncoupled system and (5.2) is not permanent. Hence (5.2) is C^1 robustly permanent if and only if $K_i > \frac{m_i}{e_i - b_i m_i} > 0$ for $i = 1, 2$.

To deduce that (5.1) is permanent for sufficiently small $\epsilon \ge 0$ when (5.2) is C^1 robustly permanent, we invoke Proposition 5.1. To this end, let $\alpha = \min\{\frac{a_1}{e_1}, \frac{a_2}{e_2}, \frac{d_1}{f_1}, \frac{d_2}{f_2}\}$ and $\beta = \min\{m_1, m_2\}$. Define $S : \mathbf{R}^4_+ \to \mathbf{R}$ by $S(x_1, y_1, x_2, y_2) = x_1 + \alpha y_1 + x_2 + \alpha y_2$. Our choice of α and β imply that any solution $(x_1(t), y_1(t), x_2(t), y_2(t))$ to (5.1) with $x_i(0) \ge 0$, $y_i(0) \ge 0$, and $\epsilon \ge 0$ satisfies

$$\begin{aligned}
&\frac{d}{dt}S(x_1(t), y_1(t), x_2(t), y_2(t)) + \beta S(x_1(t), y_1(t), x_2(t), y_2(t))\\
&\le r_1 x_1(t)\left(1-\frac{x_1(t)}{K_1}\right) + r_2 x_2(t)\left(1-\frac{x_2(t)}{K_2}\right) + \beta(x_1(t)+x_2(t)) \le C,
\end{aligned}$$

where $C = \frac{(r_1+\beta)^2 K_1}{4r_1} + \frac{(r_2+\beta)^2 K_2}{4r_2}$. It follows that

$$\limsup_{t\to\infty} S(x_1(t), y_1(t), x_2(t), y_2(t)) \le \frac{C}{\beta}.$$

Therefore, the maximal compact invariant set of (5.1) restricted to $\mathbf{R}^4_+$ for any $\epsilon \geq 0$ lies in the compact set

$$V_1 = \left\{ (x_1, y_1, x_2, y_2) \in \mathbf{R}^4_+ : S(x_1, x_2, x_3, x_4) \leq \frac{C}{\beta} \right\}.$$

Choosing any compact neighborhood V_2 of V_1, Proposition 5.1 implies that whenever (5.2) is C^1 robustly permanent, there exists a $\tilde{\epsilon} > 0$ such that (5.1) is permanent for all $0 \leq \epsilon \leq \tilde{\epsilon}$. □

Acknowledgments. The author thanks Janusz Mierczyński and two anonymous referees for helpful comments and suggestions.

REFERENCES

[1] M. Begon, J. L. Harper, and C. R. Townsend, *Ecology: Individuals, Populations and Communities*, Blackwell Scientific, Boston, 1990.

[2] E. L. Berlow, *Strong effects of weak interactions in ecological communities*, Nature, 398 (1999), pp. 330–334.

[3] G. J. Butler, H. I. Freedman, and P. Waltman, *Uniformly persistent systems*, Proc. Amer. Math. Soc., 96 (1986), pp. 425–430.

[4] G. J. Butler and P. Waltman, *Persistence in dynamical systems*, J. Differential Equations, 63 (1986), pp. 255–263.

[5] H. I. Freedman and P. Waltman, *Mathematical analysis of some three species food chains*, Math. Biosci., 68 (1977), pp. 213–231.

[6] T. C. Gard, *Persistence in food chains with general interactions*, Math. Biosci., 51 (1980), pp. 165–174.

[7] M. W. Hirsch, *Differential Topology*, Grad. Texts in Math. 33, Springer-Verlag, New York, 1976.

[8] M. W. Hirsch, H. L. Smith, and X. Zhao, *Chain transitivity, attractivity, and strong repellors for semidynamical systems*, J. Dynam. Differential Equations, 13 (2001), pp. 107–131.

[9] J. Hofbauer, *A general cooperation theorem for hypercycles*, Monatsh. Math., 91 (1981), pp. 233–240.

[10] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, Cambridge, UK, 1998.

[11] R. D. Holt, J. Grover, and D. Tilman, *Simple rules for interspecific dominance in systems with exploitative and apparent competition*, Amer. Natur., 144 (1994), pp. 741–771.

[12] S. B. Hsu, S. P. Hubbell, and P. Waltman, *Competing predators*, SIAM J. Appl. Math., 35 (1978), pp. 617–625.

[13] V. Hutson, *A theorem on average Liapunov functions*, Monatsh. Math., 98 (1984), pp. 267–275.

[14] V. Hutson, *The stability under perturbations of repulsive sets*, J. Differential Equations, 76 (1988), pp. 77–90.

[15] V. Hutson and K. Schmitt, *Permanence and the dynamics of biological systems*, Math. Biosci., 111 (1992), pp. 1–71.

[16] A. Klebanoff and A. Hastings, *Chaos in a three species food chain*, J. Math. Biol., 32 (1994), pp. 427–451.

[17] Y. A. Kuznetsov and S. Rinaldi, *Remarks on food chain dynamics*, Math. Biosci., 134 (1996), pp. 1–33.

[18] R. M. May, *Stability and Complexity in Model Ecosystems*, 2nd ed., Princeton University Press, Princeton, 1975.

[19] K. McCann, A. Hastings, and G. R. Huxel, *Weak trophic interactions and the balance of nature*, Nature, 395 (1998), pp. 794–798.

[20] K. McCann and P. Yodzis, *Bifurcation structure of a three-species food chain model*, Theor. Population Biol., 48 (1995), pp. 93–125.

[21] S. A. Navarrete and B. A. Menge, *Keystone predation and interaction strength: Interactive effects of predators on their main prey*, Ecol. Monogr., 66 (1996), pp. 409–429.

[22] R. T. Paine, *Food webs: Linkage, interaction strength and community infrastructure*, J. Anim. Ecol., 49 (1980), pp. 667–685.

[23] S. L. Pimm, *The Balance of Nature?*, The University of Chicago Press, Chicago, 1991.

[24] G. A. Polis and R. D. Holt, *Intraguild predation: The dynamics of complex trophic interactions*, Trends Ecol. Evol., 7 (1992), pp. 151–154.

[25] G. A. Polis, C. A. Meyers, and R. D. Holt, *The ecology and evolution of intraguild predation: Potential competitors that eat each other*, Annu. Rev. Ecol. Syst., 20 (1989), pp. 297–330.

[26] S. J. Schreiber, *Criteria for C^r robust permanence*, J. Differential Equations, 162 (2000), pp. 400–426.

[27] P. Schuster, K. Sigmund, and R. Wolff, *Dynamical systems under constant organization* 3: *Cooperative and competitive behavior of hypercycles*, J. Differential Equations, 32 (1979), pp. 357–368.

[28] J. Vandermeer, *Loose coupling of predator-prey cycles: Entrainment, chaos and intermittency in the classic Macarthur consumer-resource equations*, Amer. Natur., 141 (1993), pp. 687–716.

[29] P. Waltman, *A brief survey of persistence*, in Delay Differential Equations and Dynamical Systems, Lecture Notes in Math. 1475, Springer-Verlag, New York, 1991, pp. 31–40.

[30] J. T. Wootton, *Estimates and tests of per capita interaction strength: Diet, abundance, and impact of intertidally foraging birds*, Ecol. Monogr., 67 (1997), pp. 45–64.

SIAM J. MATH. ANAL.
Vol. 33, No. 5, pp. 1058–1089

CRITICAL THRESHOLD AND STABILITY OF CLUSTER SOLUTIONS FOR LARGE REACTION-DIFFUSION SYSTEMS IN R^1*

JUNCHENG WEI[†] AND MATTHIAS WINTER[‡]

Abstract. We study a large reaction-diffusion system which arises in the modeling of catalytic networks and describes the emerging of cluster states. We construct single cluster solutions on the real line and then establish their stability or instability in terms of the number N of components and the connection matrix. We provide a rigorous analysis around the single cluster solutions, which is new for systems of this kind. Our results show that for $N \leq 4$ the hypercycle system is linearly stable, while for $N \geq 5$ the hypercycle system is linearly unstable.

Key words. pattern formation, stability, cluster solutions, reaction-diffusion system, catalytic network, hypercycle

AMS subject classifications. Primary, 35B35, 92C40; Secondary, 35B40

PII. S0036141000381704

1. Introduction: The model.

In this paper, we continue our study [61] on the cluster solutions for large reaction-diffusion systems. A typical example is the hypercyclical reaction-diffusion system which arises as a spatial model concerning the origin of life similar to the one introduced by Eigen and Schuster [18], [19], [20], [21]. For more background on the concept of the hypercycle, see also [35], [36]. It arises in the modeling of catalytic networks in the case that a number of RNA-like polymers ("components") catalyze the replication of each other in a cyclic way. Examples in nature include the Krebs cycle for biosynthesis in the living cell and the Bethe–Weizsäcker cycle for high rate energy production in massive stars. Eigen and Schuster argue that the hypercycle satisfies important criteria of natural selection: (1) selective stability of each component due to favorable competition with error copies, (2) cooperative behavior of the components integrated into the hypercycle, and (3) favorable competition of the hypercycle unit with other less efficient systems.

We show rigorously that this may lead to compartmentation (i.e., the build-up of spatially small and essentially closed subsystems) due to spontaneous formation of clusters (also called "spots" or "spikes").

We first study a general system of $N + 1$ equations, where N may be any fixed positive integer representing the number of components. For this general system we first prove the existence of solutions with clusters which, for the different components, have the same location and different heights.

Then we study the stability question for some particularly important examples. At this point, we should like to emphasize that we provide a rigorous analysis around cluster solutions—not around *constant states.* We also establish a threshold size for the system such that smaller systems are stable and larger ones are unstable. This type of result is new for the kind of $(N + 1)$-systems under investigation.

*Received by the editors November 29, 2000; accepted for publication (in revised form) September 18, 2001; published electronically February 14, 2002. This research was supported by an Earmarked Research Grant from RGC of Hong Kong.

http://www.siam.org/journals/sima/33-5/38170.html

[†] Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong (wei@math.cuhk.edu.hk).

[‡] Mathematisches Institut, Universität Stuttgart, D-70511 Stuttgart, Germany (winter@mathematik.uni-stuttgart.de).

Now we proceed to write down the reaction-diffusion system explicitly and define the biological terms in a mathematically rigorous way. As suggested in [8], [9] we study the following:

$$(1.1)\begin{cases} \frac{\partial X_i}{\partial t} = D_X \Delta X_i - g_X X_i + M \sum_{j=1}^N k_{ij} X_i X_j, \quad i = 1, 2, \dots, N, \quad x \in R, \\ \frac{\partial M}{\partial t} = D_M \Delta M + k_M - g_M M - LM \sum_{i,j=1}^N k_{ij} X_i X_j, \quad x \in R, \end{cases}$$

where N is the number of different polymer species, X_i denotes the concentration of the polymers, and M is the concentration of activated monomers. The replication of each polymer X_i is catalyzed by each X_j at a nonnegative rate constant k_{ij}. Linear (noncatalytic) growth terms are neglected. The activated monomers are produced at a constant rate k_M; g_X and g_M are decay rate constants. L is the number of monomers in each polymer, and D_X and D_M are constant diffusion coefficients.

A typical example of the matrix k_{ij} is a hypercyclical $N \times N$ matrix, namely

$$(1.2) \qquad (k_{ij}^{hyper}) = \begin{pmatrix} 0 & 0 & 0 & ... & k_0 \\ k_0 & 0 & 0 & ... & 0 \\ 0 & k_0 & 0 & ... & 0 \\ ... & ... & ... & ... & 0 \\ 0 & 0 & ... & k_0 & 0 \end{pmatrix}_{N \times N}, \quad k_0 > 0.$$

The system (1.1) with the matrix (k_{ij}^{hyper}) is called "elementary hypercycle" by Eigen and Schuster [21] as the polymers interact in pairs only. There are more complex hypercycles if the polymers interact in triples, quadruples, etc. However, more complex hypercycles are likely to be of less importance for an efficient start of evolution than elementary hypercycles since they are more difficult to form in the first place.

While Eigen and Schuster [21] use an assumption of constant organization, meaning that the total sum of all polymer concentrations is kept constant, in system (1.1) another mechanism for bounding the polymer concentrations is present: Since each polymer consists of L monomers, the polymer concentrations are bounded by the limited supply of activated monomers. This is a nonlocal coupling in contrast to the local coupling in the model of Eigen and Schuster.

We pose the problem in one-dimensional space, which, on the one hand, allows a rigorous analysis and, on the other hand, is relevant if the early biochemical reactions take place in very thin lines, like, for example, on the edges of rocks.

A cluster may loosely be defined as a region of high concentrations X_i of the polymers and low concentration M of the monomer, as monomers are consumed by the replication of polymers. A rigorous definition of a cluster is given by the solution in the existence theorem (Theorem 2.1).

In this paper, we study the existence and stability of a single-cluster solution in R^1. Let us first reduce the system (1.1) to standard form. Dividing by g_X and g_M, respectively, gives

$$\frac{1}{g_X} \partial_t X_i = \frac{D_X}{g_X} \Delta X_i - X_i + \frac{M}{g_X} \sum_{j=1}^N k_{ij} X_i X_j,$$

$$\frac{1}{g_M} \partial_t M = \frac{D_M}{g_M} \Delta M + \frac{k_M}{g_M} - M - \frac{LM}{g_M} \sum_{ji,=1}^N k_{ij} X_i X_j.$$

Rescaling $M = (k_M/g_M)\hat{M}$, $X_i = \sqrt{g_M/L}\hat{X}_i$, we get

$$\frac{1}{g_X}\partial_t \hat{X}_i = \frac{D_X}{g_X}\Delta \hat{X}_i - \hat{X}_i + \frac{1}{g_X}\frac{k_M}{g_M}\hat{M}\sqrt{\frac{g_M}{L}}\sum_{j=1}^N k_{ij}\hat{X}_i\hat{X}_j,$$

$$\frac{1}{g_M}\partial_t \hat{M} = \frac{D_M}{g_M}\Delta \hat{M} + 1 - \hat{M} - \hat{M}\sum_{i,j=1}^N k_{ij}\hat{X}_i\hat{X}_j.$$

Rescaling space variables x and time variable t:

$$x = \sqrt{\frac{D_M}{g_M}}\hat{x}, \quad t = \frac{1}{g_X}\hat{t},$$

renaming constants:

$$A = \frac{k_M}{g_X g_M}\sqrt{\frac{g_M}{L}}, \quad \epsilon^2 = \frac{D_X}{D_M}\frac{g_M}{g_X}, \quad \tau = \frac{g_X}{g_M},$$

and dropping the hats, we finally arrive at the following standard form:

$$\left\{ \begin{array}{c} \partial_t X_i = \epsilon^2 \Delta X_i - X_i + AM\sum_{j=1}^N k_{ij}X_iX_j, \\ \tau\partial_t M = \Delta M + 1 - M - M\sum_{i,j=1}^N k_{ij}X_iX_j. \end{array} \right. \tag{1.3}$$

We shall study (1.3) on the real line R for $\epsilon > 0$ small. Different choices of A and τ might distinguish between stability and instability. Therefore, we will treat them as parameters. We look for solutions of (1.3) which are even:

$$X_i = X_i(|x|) \in H^1(R), \quad i = 1, \dots, N,$$

$$1 - M = 1 - M(|x|) \in H^1(R).$$

The stationary equation corresponding to (1.3) becomes

$$\left\{ \begin{array}{c} \epsilon^2 \Delta X_i - X_i + AM\sum_{j=1}^N k_{ij}X_iX_j = 0, \quad i = 1, \dots, N, \\ \Delta M + 1 - M - M\sum_{i,j=1}^N k_{ij}X_iX_j = 0, \\ X_i(|x|) > 0, \, 0 < M(|x|) < 1, \quad x \in R. \end{array} \right. \tag{1.4}$$

From now on, we shall concentrate on (1.3) and (1.4).

2. Main results: Existence and stability. Now we state our main results of this paper. First we construct cluster solutions to (1.4). To this end, we need to introduce some assumptions and notation.

Let w be the unique solution [26], [31] of the following problem:

$$\left\{ \begin{array}{l} \Delta w - w + w^2 = 0, \quad w > 0 \text{ in } R, \\ w(0) = \max_{y\in R} w(y), \quad w(y) \to 0 \text{ as } |y| \to +\infty. \end{array} \right. \tag{2.1}$$

Since (2.1) is an ODE, we can write w explicitly:

$$w(y) = \frac{3}{2\cosh^2 \frac{y}{2}}. \tag{2.2}$$

Now we state the existence result. In fact, this is quite easy. We search for solutions of the following type:

$$X_i = \xi_i X_0, \quad \xi_i > 0, \quad i = 1, \ldots, N, \tag{2.3}$$

where ξ_i are positive constants which satisfy

$$\sum_{j=1}^{N} k_{ij}\xi_j = 1, \quad i = 1, \ldots, N. \tag{2.4}$$

Our first assumption is that

(H1) there exists a unique solution $(\xi_1, \ldots, \xi_N)$ of (2.4).

Suppose (H1) holds true. Substituting (2.3) into (1.4), we see that (X_0, M) must satisfy

$$\begin{cases} \epsilon^2 \Delta X_0 - X_0 + AMX_0^2 = 0 \text{ in } R, \\ \Delta M + 1 - M - M(\sum_{i=1}^{N} \xi_i) X_0^2 = 0 \text{ in } R. \end{cases} \tag{2.5}$$

In the case in which $N = 1$, (2.5) becomes the standard Gray–Scott model [23], [24], [58]. The existence of single-pulse solutions for the Gray–Scott model in one dimension has been studied in [14] and in two dimensions in [58].

Following the same proof as in Theorem 2.1 of [58], we define

$$L = L(A, \epsilon) := \frac{1}{2A^2 \sum_{i=1}^{N} \xi_i} \epsilon \int_R (w(y))^2 dy. \tag{2.6}$$

If $0 < L < \frac{1}{4}$, then the following equation has two solutions:

$$\eta(1 - \eta) = L. \tag{2.7}$$

We denote the smaller one by η^s, where $0 < \eta^s < \frac{1}{2}$, and the larger one by η^l, where $1 > \eta^l > \frac{1}{2}$.

Now we have the following theorem.

Theorem 2.1. *Suppose that* (H1) *holds.*

Assume that

$$\epsilon << 1 \tag{2.8}$$

and

$$\epsilon << L < \frac{1}{4} - \delta_0; \tag{2.9}$$

more precisely, for $L = L(A, \epsilon)$ there are positive numbers δ_0, δ_1, and ϵ_0 such that, for all ϵ and A with $0 < \epsilon < \epsilon_0$, we have $L < \frac{1}{4} - \delta_0$ and $\epsilon / L(A, \epsilon) < \delta_1$.

Then (1.4) *admits two "single-cluster" solutions $(X_\epsilon^s, M_\epsilon^s) = (X_{\epsilon,1}^s, \ldots, X_{\epsilon,N}^s, M_\epsilon^s)$ and $(X_\epsilon^l, M_\epsilon^l) = (X_{\epsilon,1}^l \ldots, X_{\epsilon,N}^l, M_\epsilon^l)$ with the following properties:*

(1) *All components are even functions.*

(2) $X_{\epsilon,i}^s = \frac{\xi_i}{AM_\epsilon^s(0)}(1 + o(1)) w(\frac{|x|}{\epsilon})$, $X_{\epsilon,i}^l = \frac{\xi_i}{AM_\epsilon^l(0)}(1 + o(1)) w(\frac{|x|}{\epsilon})$, $i = 1, \ldots, N$, *where w is the unique solution of* (2.1).

(3) *$M_\epsilon^s(x) \to 1$ $M_\epsilon^l(x) \to 1$ for all $x \neq 0$ and $M_\epsilon^s(0)$, $M_\epsilon^l(0)$ satisfy*

$$\begin{gathered} M_\epsilon^s(0) \sim \eta^s, \quad M_\epsilon^l(0) \sim \eta^l, \\ 0 < M_\epsilon^s(0) < M_\epsilon^l(0) < 1. \end{gathered} \tag{2.10}$$

(4) *There exist $a > 0, b > 0$ such that*

$$\begin{gathered} 0 < 1 - M_\epsilon^s(x) \le Ce^{-a|x|}, \quad 0 < 1 - M_\epsilon^l(x) \le Ce^{-a|x|}, \\ 0 < X_{\epsilon,i}^s(x) \le C(AM_\epsilon^s(0))^{-1} e^{-b\frac{|x|}{\epsilon}}, \quad 0 < X_{\epsilon,i}^l(x) \le C(AM_\epsilon^l(0))^{-1} e^{-b\frac{|x|}{\epsilon}}. \end{gathered} \tag{2.11}$$

Finally, if ϵ is small enough and $L > \frac{1}{4} + \delta_0$ (in the same sense as in (2.9)), then there are no single-cluster solutions.

The proof of Theorem 2.1 is exactly the same as the proof of Theorem 2.1 of [58] or Theorem 1.1 of [61]. We omit the details here.

The main goal of this paper is to study the stability and instability of the cluster solution constructed in Theorem 2.1. To this end, we first linearize (1.3) around $(X_\epsilon^s, M_\epsilon^s)$ or $(X_\epsilon^l, M_\epsilon^l)$, respectively. From now on, we omit the superscripts s or l where this is possible without confusing the reader. The linearized operator is as follows:

$$\mathcal{L}_\epsilon \begin{pmatrix} \phi_{\epsilon,i} \\ \\ \psi_\epsilon \end{pmatrix} = \begin{pmatrix} \epsilon^2 \Delta\phi_{\epsilon,i} - \phi_{\epsilon,i} + AM_\epsilon \sum_{j=1}^N k_{ij}(\phi_{\epsilon,j} X_{\epsilon,i} + X_{\epsilon,j}\phi_{\epsilon,i}) \\ +A\psi_\epsilon \sum_{j=1}^N k_{ij} X_{\epsilon,i} X_{\epsilon,j}, \\ \Delta\psi_\epsilon - \psi_\epsilon - \psi_\epsilon \sum_{i,j=1}^N k_{ij} X_{\epsilon,i} X_{\epsilon,j} \\ -M_\epsilon \sum_{i,j=1}^N k_{ij}(\phi_{\epsilon,j} X_{\epsilon,i} + \phi_{\epsilon,i} X_{\epsilon,j}) \end{pmatrix}, \tag{2.12}$$

where $i = 1, \ldots, N$. The eigenvalue problem becomes

$$\mathcal{L}_\epsilon \begin{pmatrix} \phi_{\epsilon,i} \\ \psi_\epsilon \end{pmatrix} = \begin{pmatrix} \lambda_\epsilon \phi_{\epsilon,i} \\ \tau\lambda_\epsilon \psi_\epsilon \end{pmatrix}, \quad i = 1, \ldots, N. \tag{2.13}$$

We assume that the domain of $\mathcal{L}_\epsilon$ is $(H^2(R))^N$ and $\lambda_\epsilon \in \mathcal{C}$ is the set of complex numbers.

Certainly 0 is an eigenvalue of $\mathcal{L}_\epsilon$. We say that a cluster solution is *linearly stable* if the spectrum $\sigma(\mathcal{L}_\epsilon)$ of $\mathcal{L}_\epsilon$ (except for 0) lies in a left half-plane $\{\lambda \in \mathcal{C} : \text{ Re }(\lambda) < -c_0\}$, where $c_0 > 0$, and that 0 is a simple eigenvalue. A cluster solution is called *linearly unstable* if there exists an eigenvalue λ_ϵ of $\mathcal{L}_\epsilon$ with Re $(\lambda_\epsilon) > 0$. (From now on, we use the terms linearly stable and linearly unstable as defined above.)

Before we state our results on the stability, we introduce two more assumptions on the connection matrix (k_{ij}).

The second assumption is the following:

$$\sum_{i=1}^N k_{ij}\xi_i = 1, \quad j = 1, \ldots, N, \tag{H2}$$

where ξ_j is given (2.4).

Note that assumption (H2) imposes a certain symmetry on the connection matrix (k_{ij}).

The last assumption concerns the following eigenvalue problem:

$$\begin{cases} \Delta\phi - \phi + \mu w\phi = 0, \\ \phi \in H^1(R). \end{cases} \tag{EVP}$$

By Lemma 4.1 of [51], (EVP) admits the following set of eigenvalues:

$$\mu_1 = 1,\ \mu_2 = 2,\ 2 < \mu_3 \leq \mu_4 \leq \cdots . \tag{2.14}$$

(In fact, we have the following explicit values of μ_n (see Appendix A):

$$\mu_n = \frac{(1+n)(2+n)}{6}, \quad n = 1, 2, 3, \dots, .\Big) \tag{2.15}$$

Put

$$\mathcal{B} = (b_{ij}), \quad \text{where } b_{ij} = (\xi_i k_{ij}). \tag{2.16}$$

Observe that by (2.4) and (H1) the matrix $\mathcal{B}$ has an eigenvalue 1, and the associated eigenvector is $\xi = (\xi_1, \xi_2, \dots, \xi_N)^\tau$; i.e., we have $\mathcal{B}\xi = \xi$.

We take the Jordan decomposition of $\mathcal{B}$,

$$\mathcal{B} = \mathcal{P}\mathcal{D}\mathcal{P}^{-1}, \tag{2.17}$$

where $\mathcal{P}$ is an invertible matrix and $\mathcal{D}$ is the Jordan form. Namely, we have

$$b_{ij} = \sum_{k,l=1}^{N} p_{ik} d_{kl} p_{lj}^{-1},$$

where d_{kl} has Jordan form (i.e., it is composed of Jordan blocks

$$\begin{pmatrix} \sigma_k & 1 & 0 & \cdots & 0 \\ 0 & \sigma_k & 1 & \cdots & 0 \\ 0 & 0 & \sigma_k & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 1 \\ 0 & 0 & 0 & \cdots & \sigma_k \end{pmatrix}$$

with eigenvalues $\sigma_k \in \mathcal{C}$) and $\sum_{k=1}^{N} p_{ik} p_{kj} = \delta_{ij}$.

We now assume that

$$\begin{cases} [1 + \operatorname{spec}(\mathcal{B})] \cap \operatorname{spec}(\mathrm{EVP}) = \{2\}, \\ \ \ 1 \text{ is a simple eigenvalue of } \mathcal{B}. \end{cases} \tag{H3}$$

Assumption (H3) means the following: Let us denote the eigenvalues of $\mathcal{B}$ by

$$\sigma_1 = 1, \sigma_2, \dots, \sigma_N, \tag{2.18}$$

where σ_j may be complex. Then assumption (H3) is equivalent to

$$\sigma_j \neq \frac{(1+n)(2+n)}{6} - 1 \quad \text{for } j \geq 2, n = 1, 2, \dots . \tag{2.19}$$

Since $\xi = (\xi_1, \dots, \xi_N)^\tau$ is an eigenvector of $\mathcal{B}$ with eigenvalue 1, by assumption (H3) we may assume that

$$\mathcal{P} = (\mathbf{p}_1, \dots, \mathbf{p}_N), \quad \mathbf{p}_1 = \frac{1}{\|\xi\|}\xi, \quad \|\xi\| = \sqrt{\sum_{i=1}^{N} \xi_i^2}. \tag{2.20}$$

The following is our main result on stability.

THEOREM 2.2. *Suppose that the matrix* (k_{ij}) *satisfies* (H1), (H2), *and* (H3).

Assume that

$$\epsilon << 1, \quad \epsilon << L < \frac{1}{4} - \delta_0, \tag{2.21}$$

in the same sense as in (2.9).

Let $(X^s_\epsilon, M^s_\epsilon)$ *and* $(X^l_\epsilon, X^l_\epsilon)$ *be the solutions constructed in Theorem* 2.1.

Let $\sigma = \sigma_R + i\sigma_I$ *be an eigenvalue of* $\mathcal{B}$*, and let*

$$f(\sigma) := (12\sigma_R + 5)^2(3\sigma_R^2 + 2\sigma_R) - 3\sigma_I^2. \tag{2.22}$$

Then we have the following:

(1) *(Stability) Suppose that* $0 \leq \tau < \tau_0$*, where* $\tau_0 > 0$ *may be chosen independent of* ϵ*. Assume that for all eigenvalues* σ *of* $\mathcal{B}$ *with* $\sigma \neq 1$ *and* $\sigma_R > 0$*, we have* $f(\sigma) < 0$*. Then* $(X^s_\epsilon, M^s_\epsilon)$ *is linearly stable.*

(2) *(Instability) Assume that there exists an eigenvalue* σ *of* $\mathcal{B}$ *with* $\sigma \neq 1$ *and* $\sigma_R > 0$ *such that* $f(\sigma) > 0$*. Then* $(X^s_\epsilon, M^s_\epsilon)$ *is linearly unstable for all* $\tau > 0$.

(3) *(Instability)* $(X^l_\epsilon, M^l_\epsilon)$ *is linearly unstable for all* $\tau > 0$.

Theorem 2.2 applies to many matrices. In section 4, we shall apply Theorem 2.2 to some specific examples which include the N-hypercycle case, $(k_{ij}) = (k_{ij}^{hyper})$, where (k_{ij}^{hyper}) is given by (1.2). In this case, we have the following theorem.

THEOREM 2.3. *Consider the hypercycle case, i.e., let* (k_{ij}) *be given in* (1.2).

Assume that (2.21) *holds. Let* $(X^s_\epsilon, M^s_\epsilon)$ *and* $(X^l_\epsilon, X^l_\epsilon)$ *be the solutions constructed in Theorem* 2.1.

Then we have the following:

(1) *(Stability) Assume that* $N \leq 4$ *and* $0 < \tau < \tau_0$ *for some small* $\tau_0 > 0$ *which is independent of* ϵ*. Then* $(X^s_\epsilon, M^s_\epsilon)$ *is linearly stable.*

(2) *(Instability) Assume that* $N > 4$*. Then* $(X^s_\epsilon, M^s_\epsilon)$ *is linearly unstable for all* $\tau > 0$.

(3) *(Instability)* $(X^l_\epsilon, M^l_\epsilon)$ *is linearly unstable for all* $\tau > 0$.

The proof of Theorem 2.3 is based on Theorem 2.2 and will be given in section 4.

Some remarks on the stability results—Theorems 2.2 and 2.3—are in order.

Remarks.

(1) For existence (Theorem 2.1), only assumption (H1) is needed. For the stability results (Theorem 2.2), we need all three assumptions (H1)–(H3). Conditions (H2) and (H3) are needed in the reduction process (section 6, Lemma 6.4) and in the study of the vectorial nonlocal eigenvalue problem (NLEP) (section 7). These conditions enable us to *decouple* the system. It is an interesting open problem to study the case when assumptions (H2) and (H3) are dropped.

Note also that it is allowed that $\xi_i \neq \xi_j$ for $i \neq j$. So we may have clusters with *different heights.*

(2) In (1) of Theorem 2.2, we have assumed that τ is small. In the case in which τ is large, we can show that the stability of $(X^s_\epsilon, M^s_\epsilon)$ can be reduced to the study of an algebraic equation (section 5). More precisely, one can use hypergeometric functions and generalized hypergeometric functions to reduce the stability of the NLEP given in (5.2) to the algebraic equation which is given in Lemma 5.4 and derived in Appendix B.

(3) The threshold of stability at $N = 4$ for the hypercycle system (Theorem 2.3) has far-reaching consequences for biological applications. It implies that the underlying biological system can be stable only if it does not have too many constituents. This shows that prebiotic evolution might fail if the system becomes too large.

This is qualitatively the same result as has been established by the authors in the two-dimensional system. However, in two dimensions we were not able to establish the exact threshold [61].

Knowing the exact threshold size for stability is also important in verifying the validity of our model by experiments: Now the question can be studied of whether the thresholds given by theory and the one determined by experiments are the same. Furthermore, the agreement between theoretical values and numerically calculated ones for related models plays an important role in deciding which model to choose. (We refer to the works quoted at the end of the introduction for related numerical investigations, in particular in [7], where, among others, multicluster states in one space dimension have been computed numerically.)

Our critical threshold is in correspondence with the result of Eigen and Schuster [21] that the *constant* nontrivial steady state for the hypercycle is stable if and only if $N \leq 4$.

To see quickly how the magic number 4 comes into play, we have to study an eigenvalue problem with complex coefficients:

$$\Delta\phi - \phi + (1 + e^{\sqrt{-1}\theta})w\phi = \lambda\phi, \quad \phi \in H^2(R), \tag{2.23}$$

where $\theta = \frac{2\pi}{N}$. By using hypergeometric functions, we show (in section 5) that problem (2.23) is stable if and only if $\theta > \theta^h \sim \arccos(0.0455)$. Substituting the expression for $\theta = \frac{2\pi}{N}$, we see that $N \leq 4$.

Let us conclude this section by mentioning some related results.

In [8] the parameter dependence of the stability of clusters and spirals against parasites (i.e., rival polymers which receive catalytic support from the hypercycle but do not contribute to the catalysis of any other polymer) is studied numerically. Mathematically speaking, the occurrence of a parasite means that there exists $i_0 \in \{1, 2, \ldots, N\}$ such that $k_{i_0,j} > 0$ for some $j \neq i_0$ but $k_{j,i_0} = 0$ for all j. A parasite may or may not destroy the hypercycle depending on the rate constants. In [9] clusters (for $N = 5$) are established numerically for the elementary N-hypercycle system in two space dimensions.

It is known numerically [8], [9] that parasites may destroy stable cluster states. Our results complement the picture by the rigorously proved fact that even pure cluster states may turn unstable if they become too large. This implies that the hypercycle, although it has some very preferable properties (see the beginning of the introduction), it has an inherent instability behavior which may act as an obstruction to the evolution of large biological systems.

In [7], for a closely related reaction-diffusion model in one and two space dimensions, the dependence of various properties of cluster states on diffusivities is shown numerically, including the cluster size, their shape, and the distance between different clusters.

The effect of faulty replication on the hypercycle has been studied by an analysis of the geometry of bifurcations around steady states and numerical computations in the framework of an ODE reaction model [1].

For a cellular automata model it was shown numerically that a spiral wave structure may be stable against parasites [5]. The chaotic dynamics for this type of model has been investigated numerically in [34], [46].

There are many recent results on the Gray–Scott model which we would like to recall here. In [14], by using the Mel'nikov method, Doelman, Kaper, and Zegeling constructed single- and multiple-pulse solutions for (1.1) in the one-dimensional case with $D_M = 1, D_X = \delta^2 << 1$, where $X_i = X$. In their paper [14], it is assumed that $k_M = g_M = \delta^2, g_X = \delta^{2\alpha/3}$, $k_{11} = 1$, $L = 1$, where $\alpha \in [0, \frac{3}{2})$. In this case, they showed that $M = O(\delta^\alpha), X = O(\delta^{-\frac{\alpha}{3}})$. Later the stability of single- and multiple-pulse solutions in one dimension are obtained in [12], [13]. (The techniques are extended to other reaction-diffusion equations in [15].) We note that, in their scaling, $\tau = \delta^{\frac{2\alpha}{3}-2}$. Their scaling is chosen in order to obtain $X = O(1), M = O(1)$. Since they choose two scaling parameters accordingly, they can achieve their goal. In our standard formulation of the system (1.3), we have only the scaling parameter A so that we cannot obtain $X = O(1), M = O(1)$. On the other hand, the homoclinic solution in their scaling corresponds exactly to our cluster solution in (1.3), which is given in Theorem 1.1. For the stability results it is important to notice that the results of the system for the general N case are much more complicated than for $N = 1$. The main reason is that the behavior for the N-system cannot be reduced to the case $N = 1$ in contrast to the existence issue, and therefore a new analysis is needed.

Some related results on the existence and stability of solutions to the Gray–Scott model in one dimension can be found in [16], [29], [30], [42], [43], [47], and [48].

In R^2 and R^3, Muratov and Osipov [37] have given some formal asymptotic analysis on the construction and stability of spiky solution. In [57], the system (1.1) for $N = 1$ is studied on the real axis in the shadow system case, namely, $D_M >> 1, D_X << 1$, and $k_M = g_M = O(1), g_X = O(1)$, $k_{11} = 1$, $L = 1$. The shadow system can be reduced to a single equation. For spike solutions for single equations as well as other systems, please see [3], [4], [11], [22], [25], [27], [28], [32], [33], [38], [39], [40], [41], [44], [45], [50], [51], [52], [53], [54], [55], [56], [59], [60], and the references therein.

In the two-dimensional case, rigorous existence and stability results on the Gray–Scott system have been established in [58]. The existence of one-spike solutions is proved. Their stability is established and rests upon the derivation and analysis of a related NLEP.

3. Outline of the proof of Theorem 2.2.

We outline the proof of Theorem 2.2, which is our main theorem. It is divided into four steps. We need to analyze the eigenvalue problem (2.12). We consider two cases: small eigenvalues ($\lambda_\epsilon = o(1)$) and large eigenvalues ($|\lambda_\epsilon| \geq C > 0$ for some positive constant $C > 0$).

Step 1 (small eigenvalue case). We show that in the small eigenvalue case, λ_ϵ must be zero, and the corresponding eigenfunction must be translations of (X_ϵ, M_ϵ). This is done in Theorem 6.1 (1).

Step 2 (large eigenvalue case). We show that in the large eigenvalue case, (2.12) can be reduced to a vectorial NLEP. This is done in Theorem 6.1 (2) and (3).

Step 3 (study of the vectorial NLEP). We show, under the assumptions (H2) and (H3), that the study of the vectorial NLEP can be decoupled to the study of two eigenvalue problems—one is a scalar eigenvalue problem but with complex coefficients, and the other one is a scalar NLEP. This is done in section 7.

Step 4 (study of two eigenvalue problems). We study the two reduced eigenvalue problems in section 5. This analysis provides the key estimates in this paper.

The structure of the paper is as follows.

In section 4, we consider the applications of Theorem 2.2. In particular, we consider several interesting matrices (k_{ij}), including the hypercycle matrix and symmetric matrices.

In section 5, we study some scalar local and nonlocal eigenvalue problems associated with w.

In section 6, we separate the eigenvalue problem into two cases: small eigenvalues and large eigenvalues. The case of large eigenvalues is then linked to a vectorial NLEP given in (6.9).

In section 7, we reduce the vectorial NLEP given in (6.9) to a local eigenvalue problem with complex coefficients given in (5.1) and a scalar NLEP given in (5.2).

Throughout this paper, the letter C will always denote various generic constants which are independent of ϵ for ϵ sufficiently small. The notation $A \sim B$ means that $\lim_{\epsilon \to 0} \frac{A}{B} = 1$, and $A = O(B)$ is defined as $|A| \leq C|B|$ for some $C > 0$.

4. Effect of the connection matrix (k_{ij}).

In this section, we apply the stability results of Theorem 2.2 to some specific examples. We would like to point out that there are many matrices which satisfy assumptions (H1)–(H3) in Theorem 2.2.

Example 1 (proof of Theorem 2.3). For the hypercyclical network, we have

$$\xi_1 = \cdots = \xi_N = \frac{1}{k_0},$$

$$b_{ij}^{hyper} = \delta_{i,j+1} \quad \text{modulo } N.$$

The eigenvalues are $\sigma = e^{2\pi j\sqrt{-1}/N}$, $j = 1, \ldots, N$, and they are all simple. In this case, it is easy to see that (H1)—(H3) are satisfied. By Theorem 2.2, we just need to find the zeroes of the following function:

$$f(\sigma) := (12\sigma_R + 5)^2(3\sigma_R^2 + 2\sigma_R) - 3\sigma_I^2, \quad \sigma_R^2 + \sigma_I^2 = 1, \quad 0 < \sigma_R < 1. \tag{4.1}$$

It is easy to check that the solution to (4.1) is

$$\sigma_R^0 = 0.0455\ldots.$$

Note that $\cos(\frac{2\pi}{5}) > \sigma_R^0$.

By Theorem 2.2 (1), we obtain the stability of the small cluster solution for $N = 1, 2, 3, 4$. By Theorem 2.2 (2), we obtain the instability of small solutions for $N \geq 5$.

We conclude that the critical threshold size for the hypercycle system is 4. When the system size exceeds 4, then a parasite appears: there is an eigenvector $c = (c_1, \ldots, c_N)^\tau$ of (k_{ij}) such that $\sum_{j=1}^N c_j X_j$ vanishes quickly.

Example 2. We consider the case when the connection matrix (k_{ij}) is symmetric, i.e.,

$$k_{ij} = k_{ji}.$$

In this case, it is easy to see that the matrix $\mathcal{B} = (k_{ij}\xi_i)$ has only real eigenvalues. Let the eigenvalues of $\mathcal{B}$ be

$$\sigma_1 = 1, \sigma_2, \ldots, \sigma_N.$$

The first eigenvalue $\sigma_1 = 1$ is guaranteed by (2.4).

Assumption (H2) is satisfied if we further assume that $\xi_1 = \cdots = \xi_N$.

Assumption (H3) says that

$$\sigma_j \neq \frac{(1+k)(2+k)}{6} - 1, \quad j = 2, \ldots, N, \quad k = 1, 2, \ldots. \tag{4.2}$$

Since $f(\sigma) > 0$ if $\sigma = \sigma_R > 0$, Theorem 2.2 shows that $(X_\epsilon^s, M_\epsilon^s)$ is linearly stable if

$$\sigma_j < 0, \quad j = 2, \ldots, N. \tag{4.3}$$

On the other hand, if there exists $\sigma_j > 0$ for some $j \geq 2$, we have instability. (Assumption (H3) implies that $\sigma_j \neq 0$.)

Example 3. For the (cyclical) bidiagonal matrix

$$(k_{ij}) = k_0 \begin{pmatrix} 1-\alpha & \alpha & 0 & \ldots & 0 \\ 0 & 1-\alpha & \alpha & \ldots & 0 \\ 0 & 0 & 1-\alpha & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \alpha \\ \alpha & 0 & \ldots & 0 & 1-\alpha \end{pmatrix}_{N \times N}, \quad k_0 > 0,$$

with $0 \leq \alpha < 1$, it is easy to see that conditions (H1)–(H3) are satisfied. In this case, $\xi_1 = \cdots = \xi_N = \frac{1}{k_0}$. The eigenvalues are computed as $\sigma = 1 - \alpha(1 - e^{2\pi j\sqrt{-1}/N})$, $j = 1, \ldots, N$, and are all simple.

We substitute σ into the polynomial and compute the critical threshold $N_{critical}$. It turns out that $N_{critical}$ depends on the value of α: $N_{critical}$ will increase of the order α as α increases. The following is a table of $N_{critical}$ for small α:

α	$N_{critical}$
0.5	3
1	4
1.5	5
2	6

From all of the previous examples, we see as a general trend that, if the system is not dominated too much by diagonal terms, we have stability. Otherwise, a parasite emerges. This means that cooperative behavior and not self-enhancement is needed to stabilize the cluster.

We point to the last example, where the stability is especially strong if the parameter α gets large. In the case in which $\alpha > 1$ (which means that the diagonal becomes negative and the off-diagonal elements are positive and bigger than the diagonal), this describes self-inhibition coupled with cooperative enhancement and leads to particularly good stability.

5. Two eigenvalue problems. In this section, we study two eigenvalue problems. The first is a local eigenvalue problem with complex coefficients

$$\begin{cases} \Delta\phi - \phi + w\phi + \sigma w\phi = \lambda\phi, \\ \sigma = \sigma_R + i\sigma_I = |\sigma|e^{i\theta}, |\sigma| > 0, \quad \theta \in (-\pi, \pi], \quad \phi \in H^1(R), \end{cases} \tag{5.1}$$

where w is defined by (2.1).

The second is a scalar NLEP

$$\Delta\phi - \phi + 2w\phi - \frac{2(1-\eta)}{\eta\sqrt{1+\tau\lambda} + 1 - \eta} \frac{\int_R w\phi}{\int_R w^2} w^2 = \lambda\phi, \quad \phi \in H^2(R), \tag{5.2}$$

where

$$0<\eta<1,\ \tau\geq 0,\ \lambda\in\mathcal{C},\ \lambda=\lambda_R+i\lambda_I,\ \lambda_R\geq 0,$$

and we take the principal branch for $\sqrt{1+\tau\lambda}$.

The analysis presented in this section provides the key estimates for this paper.

To study (5.1) and (5.2), we first collect some important properties associated with the function w.

LEMMA 5.1. (1) *The linear operator*

$$\begin{cases} L_0\phi := \Delta\phi-\phi+2w\phi, \\ \qquad \phi\in H^1(R), \end{cases}$$

has the kernel

$$\text{Ker }(L_0)=\text{span }\left\{w^{'}(y)\right\}.$$

(2) *The eigenvalue problem*

$$(EVP) \qquad \begin{cases} \Delta\phi-\phi+\mu w\phi=0, \\ \qquad \phi\in H^1(R), \end{cases}$$

admits the following set of eigenvalues:

$$\mu_1=1,\ v_1=\text{span }\{w\},$$

$$\mu_2=2,\ v_2=\text{Ker }(L_0),$$

$$\mu_n=\frac{(1+n)(2+n)}{6}>2 \textit{ for } n\geq 3.$$

(3) *If $\mu_R>0$, then the eigenvalue problem*

$$\begin{cases} \Delta\phi-\phi+w\phi+\mu_R w\phi=\lambda\phi, \\ \qquad \mu_R>0,\ \phi\in H^1(R), \end{cases}$$

admits a positive (principal) eigenvalue λ_1 such that

$$-\lambda_1=\inf_{\phi\in H^1(R)\setminus\{0\}}\frac{\int_R(\phi')^2+\phi^2-(1+\mu_R)w\phi^2}{\int_R\phi^2}<0.$$

Moreover, when $\mu_R=1$, there is only one positive eigenvalue (which is the principal one).

(4) *Let ϕ (complex-valued) satisfy the following eigenvalue problem:*

$$\begin{cases} \Delta\phi-\phi+w\phi+\sigma w\phi=\lambda\phi, \\ \text{Re }(\sigma)\leq 0, \quad \phi\in H^1(R), \quad \lambda\neq 0. \end{cases}$$

Then

$$\text{Re }(\lambda)\leq -c_0<0.$$

Proof. The proof will be given in Appendix A. The proof of (2) follows the method of Lemma 5.2. Some of the results have been proved in previous work. For the convenience of the reader we recall the proofs of (3) and (4). □

We are ready to study the first eigenvalue problem (5.1). By symmetry, we may assume that $\theta \in [0, \frac{\pi}{2}]$. We consider θ as a parameter. By Lemma 5.1 (3) and a perturbation argument, for $|\theta|$ near 0, there is an unstable eigenvalue λ for (5.1), i.e., $\lambda = \lambda_R + i\lambda_I$, where $\lambda_R > 0$. On the other hand, by Lemma 5.1 (4), for $|\theta| \geq \frac{\pi}{2}$, (5.1) has only stable eigenvalues, i.e., $\lambda = \lambda_R + i\lambda_I$, where $\lambda_R < 0$. Now, if we vary θ, then there must be a point $\theta^h \in (0, \frac{\pi}{2})$ such that, for $\theta = \theta^h$, (5.1) has a Hopf bifurcation, i.e., there is an eigenvalue $\lambda = \sqrt{-1}\lambda_I$. Let us now compute θ^h. It turns out that, unlike in the two-dimensional case [61], we can now obtain the exact value for θ^h in one domension.

LEMMA 5.2. *Let ϕ (complex-valued) satisfy the eigenvalue problem* (5.1) *with $\sigma = \sigma_R + \sqrt{-1}\sigma_I, \sigma_R > 0$. Then the following hold.*

(1) *If $f(\sigma) < 0$, then* (5.1) *is stable.*

(2) *If $f(\sigma) > 0$, then* (5.1) *is unstable.*

(3) *If $f(\sigma) = 0$, then there exists an eigenvalue λ with $\lambda = \sqrt{-1}\lambda_I$.*

Here $f(\sigma) := (12\sigma_R + 5)^2(3\sigma_R^2 + 2\sigma_R) - 3\sigma_I^2$.

Proof. We are looking for a Hopf bifurcation for (5.1). Therefore, we have to solve

$$\Delta\phi - \phi + (1 + \sigma)w\phi = \lambda\phi \tag{5.3}$$

with

$$\lambda = \sqrt{-1}\lambda_I$$

(i.e., the real part λ_R of λ vanishes) and

$$\sigma = \sigma_R + i\sigma_I.$$

As in [12], let

$$\gamma = \sqrt{1 + \lambda}, \quad \mu = 1 + \sigma, \quad \phi = w^\gamma F.$$

Then F satisfies

$$F^{''} + 2\gamma\frac{w^{'}}{w}F^{'} + \left(\mu - \left(\gamma + \frac{2}{3}\gamma(\gamma - 1)\right)\right) w^{p-1}F = 0. \tag{5.4}$$

Next we introduce the following new variable:

$$z = \frac{1}{2}\left(1 - \frac{w^{'}}{w}\right). \tag{5.5}$$

Then

$$\frac{w^{'}}{w} = 1 - 2z, \quad w = 6z(1 - z), \quad \frac{dz}{dx} = z(1 - z).$$

This yields the following equation for F as a function of z:

$$z(1 - z)F^{''} + (c - (a + b + 1)z)F^{'} - abF = 0, \tag{5.6}$$

where

$$a+b+1=2+4\gamma, \quad ab=2(2\gamma(\gamma-1)-3(\mu-\gamma)), \quad c=1+2\gamma. \tag{5.7}$$

The solutions to (5.6) are standard hypergeometric functions. See [49] for more details. Now there are two solutions to (5.6):

$$F(a,b;c;z), \quad z^{1-c}F(a-c+1,b-c+1;2-c;z).$$

By our construction F is regular at $z=0$. At $z=1$, $F(a,b;c;z)$ has a singularity

$$\lim_{z\to 1}(1-z)^{-(c-a-b)}F(a,b;c;z)=\frac{\Gamma(c)\Gamma(a+b-c)}{\Gamma(a)\Gamma(b)},$$

where $c-a-b=-2\gamma<0$. Note that since $\gamma=\sqrt{1+\sqrt{-1}\lambda_I}$, the real part of γ is positive. So a solution that is regular at both $z=0$ and $z=1$ can exist only if $\Gamma(x)$ has a pole at a or b, respectively. In other words, $a,b=0,-1,-2,\ldots$.

From (5.7), we compute that

$$a=2\gamma-\alpha \text{ or } b=2\gamma-\alpha,$$

where α satisfies

$$\alpha^2+\alpha-6\mu=0. \tag{5.8}$$

By symmetry we may assume that $a=2\gamma-\alpha=-l,\ l\geq 0$, and $\alpha=\alpha_R+\sqrt{-1}\alpha_I$. So we have to solve the system

$$\begin{cases} \alpha_R^2+\alpha_R-\alpha_I^2-6(1+\sigma_R)=0, \\ 2\gamma=\alpha-l, \quad l=0,1,2,\ldots. \end{cases} \tag{5.9}$$

Since we take the principal branch for $\gamma=\sqrt{1+\sqrt{-1}\lambda_I}$, it follows that

$$\alpha_R>l.$$

Moreover, we have

$$4=(\alpha_R-l)^2-\alpha_I^2,$$

which implies that

$$\alpha_R\geq l+2. \tag{5.10}$$

On the other hand, we have

$$4=(\alpha_R-l)^2-\alpha_I^2=\alpha_R^2-\alpha_I^2-2l\alpha_R+l^2$$

$$=-(2l+1)\alpha_R+l^2+6(1+\sigma_R).$$

So we obtain

$$\alpha_R=\frac{1}{2l+1}(l^2+2+6\sigma_R).$$

By (5.10), we have

$$\frac{1}{2l+1}(l^2+2+6\sigma_R) \geq l+2,$$

which is impossible unless $l=0$ or $l=1$. For $l=1$ we just recover the case $\lambda=0$ with the eigenfunction w' given by Lemma 5.1 (1). This clearly does not correspond to a Hopf bifurcation.

In conclusion, for a Hopf bifurcation we must have $a=0$ or $b=0$. In this case, we have

$$\alpha_R = 2+6\sigma_R, \quad \alpha_I = \frac{6}{(2\alpha_R+1)}\sigma_I. \tag{5.11}$$

Substituting this relation into (5.9), we obtain that (σ_R, σ_I) must be a zero of the polynomial f defined by (2.22).

In summary, a Hopf bifurcation can occur only at the point (σ_R^h, σ_I^h) such that $f(\sigma)=0$.

Note that the set $\{f(\sigma)=0\}$ defines a monotone curve within the set $\{(\sigma_R,\sigma_I)|\sigma_R > 0, \sigma_I>0\}$. Since $f(0,\sigma_I)<0$ for $\sigma_I>0$ and $f(\sigma_R,0)>0$ for $\sigma_R>0$, we see that if $f(\sigma)<0$, then (5.1) is stable, and if $f(\sigma)>0$, then (5.1) is unstable. □

Next we study the scalar NLEP (5.2). First we state the following lemma.

LEMMA 5.3. *Consider the NLEP* (5.2).

(1) *Suppose that* $0 \leq \tau < \tau_0$, *where* τ_0 *is sufficiently small and* $0<\eta<\frac{1}{2}$. *Let* $\lambda_0 \neq 0$ *be an eigenvalue of* (5.2). *Then we have* $\mathrm{Re}(\lambda_0) \leq -c_1$ *for some* $c_1>0$.

(2) *Suppose that* $\tau>0$ *and* $\frac{1}{2}<\eta<1$; *then* (5.2) *admits a real eigenvalue* λ_0 *with* $\lambda_0 \geq c_2 > 0$ *for some* $c_2>0$.

Proof. (1) When $\tau=0$, we have

$$\frac{2(1-\eta)}{\eta\sqrt{1+\tau\lambda}+1-\eta} = 2(1-\eta) > 1$$

if $0<\eta<\frac{1}{2}$. By Theorem 2.1 of [57], we have that $\lambda_R < -c_1 < 0$.

To show that the same thing is true when τ is small, we have to show that if $\lambda_R \geq 0$ and $0<\tau<1$, then $|\lambda| \leq C$, where C is a generic constant (independent of τ). In fact, multiplying (5.2) by $\bar{\phi}$, the conjugate of ϕ, and integrating by parts, we obtain that

$$\int_R (|\nabla\phi|^2+|\phi|^2-2w|\phi|^2) = -\lambda\int_R |\phi|^2 - f(\tau\lambda)\frac{\int_R w\phi}{\int_R w^2}\int_R w^2\bar{\phi}, \tag{5.12}$$

where $f(\tau\lambda)=\frac{2(1-\eta)}{\eta\sqrt{1+\tau\lambda}+1-\eta}$. From the imaginary part of (5.12), we obtain that

$$|\lambda_I| \leq C_1|f(\tau\lambda)|,$$

where $\lambda=\lambda_R+\sqrt{-1}\lambda_I$ and C_1 is a positive constant (independent of τ). Note that the real part of $\sqrt{1+\tau\lambda}$ is positive. Hence $|f(\tau\lambda)| \leq 2$, and so $|\lambda_I| \leq 2C_1$. Taking the real part of (5.12), we obtain that $\lambda_R \leq C_2$, where C_2 is a positive constant (independent of $\tau>0$). Therefore, we have that $|\lambda|$ is uniformly bounded, and hence a perturbation argument gives the desired conclusion.

(2) Assume that $\frac{1}{2}<\eta<1$. Now we show that (5.2) admits a positive eigenvalue for all $\tau>0$.

By Lemma 5.1 (3), L_0 has only one positive eigenvalue $\lambda_1 > 0$. Consider the following function:

$$h(\alpha) = \int_R ((L_0 - \alpha)^{-1} w) w, \quad 0 < \alpha < \lambda_1. \tag{5.13}$$

It is easy to see that

$$h^{'}(\alpha) = \int_R ((L_0 - \alpha)^{-2} w) w = \int_R [(L_0 - \alpha)^{-1} w]^2 > 0,$$

and

$$\lim_{\alpha \to \lambda_1} h(\alpha) = +\infty.$$

Next we consider the following function:

$$\rho(\lambda) = \frac{\eta\sqrt{1 + \tau\lambda} + 1 - \eta}{2(1 - \eta)} - 1 - \left(\int_R w^2\right)^{-1} \lambda h(\lambda). \tag{5.14}$$

Note that

$$\rho(0) = \frac{1}{2(1 - \eta)} - 1 > 0$$

since $\frac{1}{2} < \eta < 1$. On the other hand,

$$\lim_{\lambda \to \lambda_1 -} \rho(\lambda) = -\infty.$$

Hence there must exist an $\lambda_0 \in (0, \lambda_1)$ such that $\rho(\lambda_0) = 0$. It is easy to see that this $\lambda_0 > 0$ is an eigenvalue of (5.2), which proves (2) of Lemma 5.3. □

In the general case when τ is large and $0 < \eta < \frac{1}{2}$, there are no analytic results for (5.2) available. Fortunately, we can use hypergeometric functions and generalized hypergeometric functions to reduce (5.2) to a computable problem. Such an idea has already been used in [12]. However, our transformation is different, and the eigenvalue problem becomes computable more easily. We recall that by Lemma 5.3 (1) for $\tau = 0$ all eigenvalues are stable. So, if we vary τ, we obtain either stability or a Hopf bifurcation. All we need is to compute when a Hopf bifurcation occurs.

Let us first introduce the so-called generalized Gauss function. Let $a_1, a_2, \dots, a_A$ and $b_1, b_2, \dots, b_B$ be two sequences of numbers. Consider the following series:

$$1 + \frac{a_1 a_2 \cdots a_A}{b_1 b_2 \cdots b_B} \frac{z}{1!} + \frac{(a_1 + 1)(a_2 + 1) \cdots (a_A + 1)}{(b_1 + 1)(b_2 + 1) \cdots (b_B + 1)} \frac{z^2}{2!} + \cdots \tag{5.15}$$

$$\equiv {}_A F_B \left\{ \begin{array}{ccccc} a_1, & a_2, & \dots, & a_A & ; \\ & & & & z \\ b_1, & b_2, & \dots, & b_B & ; \end{array} \right\}.$$

${}_A F_B$ is called a generalized Gauss function or a generalized hypergeometric function. For more details on such functions, we refer to [49].

Now we have the following lemma, whose proof is technical and is thus delayed to Appendix B.

Lemma 5.4. *Let $\lambda=\sqrt{-1}\lambda_I$ be an eigenvalue of* (5.2). *Then λ is a solution of the following algebraic equation:*

$$\frac{5}{6}(3-\lambda)\frac{\eta\sqrt{1+\tau\lambda}+1-\eta}{2(1-\eta)}$$

$$= {}_4F_3\left\{\begin{array}{ccccc} 1, & \frac{1}{2}, & 4, & 3 & ; \\ & & & & 1 \\ & 3-\sqrt{1+\lambda}, & 3+\sqrt{1+\lambda}, & \frac{7}{2} & ; \end{array}\right\}. \tag{5.16}$$

By Lemma 5.4, (5.2) can be solved by using Mathematica. We will not produce any numerical results here. The readers are referred to [12] for some numerical results for the case in which $N=1$.

6. Derivation of the vectorial NLEP and reduction process. In this section, we study the eigenvalue problem (2.12) and show that it can be reduced to a vectorial NLEP.

Let (X_ϵ, M_ϵ) be one of the two solutions constructed in section 2. Now we study the eigenvalue problem associated with (X_ϵ, M_ϵ). We assume that

$$\epsilon << L < \frac{1}{4}-\delta_0$$

(in the same sense as in (2.9)), where $\delta_0>0$ is a small but fixed constant, and that $0\le\tau<\tau_0$, where τ_0 is given by Lemma 5.3 and is independent of ϵ.

We need to analyze the following eigenvalue problem (letting $x=\epsilon y$):

$$\begin{cases} \Delta_y\phi_{\epsilon,i}-\phi_{\epsilon,i}+AM_\epsilon\sum_{j=1}^N k_{ij}(X_{\epsilon,j}\phi_{\epsilon,i}+\phi_{\epsilon,j}X_{\epsilon,i}) \\ +A\psi_\epsilon\sum_{j=1}^N k_{ij}X_{\epsilon,i}X_{\epsilon,j}=\lambda_\epsilon\phi_{\epsilon,i},\ y\in R, i=1,\dots,N, \\ \Delta_x\psi_\epsilon-\psi_\epsilon-\psi_\epsilon\sum_{i,j=1}^N k_{ij}X_{\epsilon,i}X_{\epsilon,j} \\ -M_\epsilon\sum_{i,j=1}^N k_{ij}(X_{\epsilon,j}\phi_{\epsilon,i}+X_{\epsilon,i}\phi_{\epsilon,j})=\tau\lambda_\epsilon\psi_\epsilon, \quad x\in R, \\ \lambda_\epsilon\in\mathcal{C}. \end{cases} \tag{6.1}$$

We assume that $(\phi_{\epsilon,1},\dots,\phi_{\epsilon,N},\psi_\epsilon)\in(H^2(R))^N\oplus H^2(R)$. Here we equip $(H^2(R))^N\oplus H^2(R)$ with the following norm:

$$\|(X,u)\|^2_{(H^2(R))^N\oplus H^2(R)}=\|X(y)\|^2_{(H^2(R))^N}+\|u(x)\|^2_{H^2(R)}.$$

Since $X_{\epsilon,i}=\xi_iX_{0,\epsilon}$, (6.1) becomes

$$\begin{cases} \Delta_y\phi_{\epsilon,i}-\phi_{\epsilon,i}+AM_\epsilon X_{0,\epsilon}\sum_{j=1}^N k_{ij}(\xi_i\phi_{\epsilon,j}+\xi_j\phi_{\epsilon,i}) \\ +A\psi_\epsilon\xi_iX_{0,\epsilon}^2=\lambda_\epsilon\phi_{\epsilon,i}, \\ \Delta\psi_\epsilon-\psi_\epsilon-\psi_\epsilon(\sum_{i=1}^N\xi_i)X_{0,\epsilon}^2 \\ -M_\epsilon\sum_{i,j=1}^N k_{ij}(\xi_i\phi_{\epsilon,j}+\xi_j\phi_{\epsilon,i})X_{0,\epsilon}=\tau\lambda_\epsilon\psi_\epsilon. \end{cases} \tag{6.2}$$

First let us formally derive the limiting eigenvalue problems.

Since $(X_{0,\epsilon}, M_\epsilon)$ satisfies (2.5), we have

$$X_{0,\epsilon}(y)\sim(AM_\epsilon(0))^{-1}(1+o(1))w(y) \tag{6.3}$$

and

$$M_\epsilon(0)(1-M_\epsilon(0)) \sim L := \frac{1}{2A^2(\sum_{i=1}^N \xi_i)}\epsilon \int_R w(y)^2 dy. \tag{6.4}$$

By the assumptions (H1) and (H2), $\sum_{j=1}^N k_{ij}\xi_j = \sum_{i=1}^N k_{ij}\xi_i = 1$, the eigenvalue problem is changed into

$$\begin{cases} \Delta_y \phi_{\epsilon,i} - \phi_{\epsilon,i} + w\phi_{\epsilon,i} + w\sum_{j=1}^N b_{ij}\phi_{\epsilon,j} \\ +\frac{1}{AM_\epsilon(0)^2}\xi_i \psi_\epsilon w^2 = \lambda_\epsilon \phi_{\epsilon,i}, \quad i=1,\ldots,N, \\ \Delta_x \psi_\epsilon - \psi_\epsilon - \frac{1}{A^2M_\epsilon(0)^2(\sum_{i=1}^N \xi_i)}\psi_\epsilon w^2 \\ -\frac{M_\epsilon}{AM_\epsilon(0)}2\sum_{j=1}^N \phi_{\epsilon,j} w = \tau\lambda_\epsilon \psi_\epsilon. \end{cases} \tag{6.5}$$

Let $\beta_\epsilon = \sqrt{1+\tau\lambda_\epsilon}$. Here we take the principal branch of $1+\tau\lambda_\epsilon$. Since we are interested only in the unstable eigenvalues of λ_ϵ (otherwise it is stable), we may assume that $\mathrm{Re}(\lambda_\epsilon) \geq -a_0$ for some small number $a_0 > 0$ so that $1+\tau a_0 > \frac{1}{2}$. Following the same proof as for (1) of Lemma 5.3 (that is, multiplying the equations for $\phi_{\epsilon,i}$ by $\bar{\phi}_{\epsilon,i}$, integrating by parts, and summing up), we see that

$$|\lambda_\epsilon| \leq C \quad \text{if } \mathrm{Re}(\lambda_\epsilon) \geq -a_0, \tag{6.6}$$

where $C > 0$ is a positive constant (independent of $\epsilon > 0$).

From the second equation in (6.5), we calculate using the fact that Green's function of

$$\Delta G(x,\xi) - \beta^2 G(x,\xi) + \delta(\xi) = 0 \quad \text{in } R$$

is

$$G(x,\xi) = \frac{1}{2\beta}e^{-\beta|x-\xi|},$$

the relation

$$\psi_\epsilon(0) = \frac{1}{2\beta_\epsilon}\int_R e^{-\beta_\epsilon |x|}$$

$$\times\left(-\frac{1}{A^2M_\epsilon(0)^2(\sum_{i=1}^N \xi_i)}\psi_\epsilon w^2 - \frac{M_\epsilon}{AM_\epsilon(0)}2\sum_{j=1}^N \phi_{\epsilon,j} w\right) dx$$

$$= \frac{1}{2\beta_\epsilon}\epsilon\Bigg[-\frac{\psi_\epsilon(0)}{A^2M_\epsilon(0)^2(\sum_{i=1}^N \xi_i)}\int_R w^2(y)\,dy$$

$$-\frac{1}{A}\sum_{j=1}^N \int_R \left(\sum_{j=1}^N \phi_j\right) w\,dy + o(\epsilon)\Bigg], \tag{6.7}$$

where

$$\phi_{\epsilon,i}(x) = \phi_i\left(\frac{x}{\epsilon}\right), \quad x = \epsilon y, \quad i = 1,\ldots,N. \tag{6.8}$$

By (6.4) and (6.7), we have

$$
\begin{aligned}
&\frac{\psi_\epsilon(0)}{AM_\epsilon^2(0)}\\
&\quad= -\frac{\epsilon}{\beta_\epsilon A^2 M_\epsilon(0)^2}\int_R \left(\sum_{j=1}^N \phi_j\right) w \left[1+\frac{\epsilon}{2\beta_\epsilon(\sum_{i=1}^N \xi_i)A^2M_\epsilon(0)^2}\int_R w^2(y)\,dy\right]^{-1}\\
&\quad= -\int_R \left(\sum_{j=1}^N \phi_j\right) w \left[\frac{\beta_\epsilon A^2 M_\epsilon(0)^2}{\epsilon}+\frac{1}{2(\sum_{i=1}^N \xi_i)}\int_R w^2(y)\,dy\right]^{-1}\\
&\quad= -\int_R \left(\sum_{j=1}^N \phi_j\right) w \left[\frac{\beta_\epsilon M_\epsilon(0)(\sum_{i=1}^N \xi_i)}{2(1-M_\epsilon(0))}\int_R w^2(y)\,dy+\frac{(\sum_{i=1}^N \xi_i)}{2}\int_R w^2(y)\,dy\right]^{-1}\\
&\quad= -\frac{2(1-M_\epsilon(0))}{1-M_\epsilon(0)+\beta_\epsilon M_\epsilon(0)}\frac{\int_R(\sum_{j=1}^N \phi_j)w}{(\sum_{i=1}^N \xi_i)\int_R w^2(y)\,dy}.
\end{aligned}
$$

Substituting this relation into the first equation in (6.5) and taking the limit $\epsilon \to 0$, we obtain the following NLEP:

$$
\begin{aligned}
(6.9)\qquad &\Delta\phi_i - \phi_i + \phi_i w + \sum_{j=1}^N b_{ij}\phi_j - \xi_i \frac{2(1-\eta)}{\eta\beta_0+1-\eta}\frac{\sum_{j=1}^N \int_R \phi_j w}{(\sum_{i=1}^N \xi_i)\int_R w^2} w^2\\
&= \lambda_0\phi_i, \phi_i \in H^2(R), i=1,\dots,N,
\end{aligned}
$$

where $\eta = \lim_{\epsilon\to 0} M_\epsilon(0), \lambda_0 = \lim_{\epsilon\to 0}\lambda_\epsilon, \beta_0 = \lim_{\epsilon\to 0}\beta_\epsilon = \sqrt{1+\tau\lambda_0}$. (Here, we have assumed that all the limits exist. Otherwise, we take a subsequence $\epsilon_n \to 0$.)

Though the derivations above are formal, we can rigorously prove the following separation of eigenvalues.

THEOREM 6.1. *Suppose that the assumptions* (H1)–(H3) *are satisfied.*

Let λ_ϵ be an eigenvalue of (6.2) *such that* $\mathrm{Re}(\lambda_\epsilon) > -a_0$.

(1) *Suppose that (for suitable sequences $\epsilon_n \to 0$) we have $\lambda_{\epsilon_n} \to 0$ as $n \to \infty$. Then for n sufficiently large, it follows that $\lambda_{\epsilon_n} = 0$ and*

$$(\phi_{\epsilon_n,1},\dots,\phi_{\epsilon_n,N},\psi_{\epsilon_n}) \in \text{span}\,\{(X'_{\epsilon_n}, M'_{\epsilon_n})\}.$$

(2) *Suppose that (for suitable sequences $\epsilon_n \to 0$) we have $\lambda_{\epsilon_n} \to \lambda_0 \neq 0$. Then λ_0 is an eigenvalue of the NLEP given in* (6.9).

(3) *Let $\lambda_0 \neq 0$ be an eigenvalue of the NLEP given in* (6.9). *Then for ϵ sufficiently small, there is an eigenvalue λ_ϵ of* (6.2) *with $\lambda_\epsilon \to \lambda_0$ as $\epsilon \to 0$.*

From Theorem 6.1 (1) and (3), we see that (6.2) is reduced to the study of the vectorial NLEP (6.9).

In the rest of this section, we prove Theorem 6.1.

Proof of Theorem 6.1. For (1), the proof is very delicate. We can proceed as in the proof of Theorem 2.2 (3) in section 6 of [58], where existence and stability of a single-cluster state for the Gray–Scott system in two dimensions are studied. First we prove the analogies of Lemmas 3.1 and 3.2 of [58] in one dimension. We begin with the following lemma.

LEMMA 6.2. *Let $g(y)$ be a function in $L^2(R^1)$ such that*

$$|g(y)| \le Ce^{-c|y|},$$

where c is a positive constant. Then we have

$$\left|\int_R (|y-z|-|z|)|g(z)|dz\right| \le C|y|, \tag{6.10}$$

where C depends on $\int_R |z||g(z)|dz$.

Proof. This follows from standard potential analysis. □

Next we study the asymptotic behavior of ψ_ϵ. We have the following lemma.

Lemma 6.3. *Let $(\phi_{\epsilon,1},\dots,\phi_{\epsilon,N},\psi_\epsilon)$ satisfy* (6.5). *Then we have*

$$\frac{1}{AM_\epsilon(0)^2}\psi_\epsilon(0) = -\frac{1-M_\epsilon(0)}{\beta_\epsilon M_\epsilon(0)+1-M_\epsilon(0)}\,\frac{2\sum_{j=1}^N\int_R\sum_{j=1}^N\phi_{\epsilon,j}w}{(\sum_{i=1}^N\xi_i)\int_R w^2}+o(1) \tag{6.11}$$

and

$$\frac{1}{AM_\epsilon(0)^2(\sum_{i=1}^N\xi_i)}(\psi_\epsilon(x)-\psi_\epsilon(0))$$

$$= O\left(\frac{2}{\epsilon(1-\sqrt{1-4L})}\left(1+\sum_{i=1}^N\|\phi_{\epsilon,i}\|_{L^2_y}\right)\left(1+\frac{|x|}{\epsilon}\right)\right), \tag{6.12}$$

where $x=\epsilon y$ and

$$\|\phi\|^2_{L^2_y}=\int_R\phi^2(y)dy.$$

Proof. Relation (6.11) follows from the representation formula. To prove (6.12), we note that by the representation formula we calculate

$$\psi_\epsilon(x)-\psi_\epsilon(0)=\frac{1}{2\beta}\int_R(e^{-\beta_\epsilon|z-x|}-e^{-\beta_\epsilon|z|})$$

$$\times\left(-\frac{\psi_\epsilon X_{0,\epsilon}^2}{(\sum_{i=1}^N\xi_i)}-2M_\epsilon\left(\sum_{j=1}^N\phi_{\epsilon,j}\right)X_{0,\epsilon}\right)dz.$$

Let $x=\epsilon y$ and $z=\epsilon\tilde z$. It is easy to see that

$$e^{-\beta_\epsilon|z-x|}-e^{-\beta_\epsilon|z|}=e^{-\beta_\epsilon\epsilon|y-\tilde z|}-e^{-\beta_\epsilon\epsilon|\tilde z|}$$

$$=-\beta_\epsilon\epsilon(|y-\tilde z|-|\tilde z|)+O(\beta_\epsilon^2\epsilon^2(|y|^2+|\tilde z|^2)).$$

Equation (6.12) now follows from Lemma 6.2. □

Finally, we need the analogue of Lemma 4.2 of [58].

Let us denote the linear operator on the left-hand side of (6.9) as $\mathcal{L}$, where $\mathcal{L}:(H^2(R))^N\to(L^2(R))^N$. Then we have the following lemma.

Lemma 6.4. *Assume that assumptions* (H1)–(H3) *hold true.*

(1) *Let ϕ be an eigenfunction of* (6.9) *with $\lambda_0=0$. Then we have*

$$\phi\in\mathcal{K}_0:=\text{span }\{w^{'}(y)\vec e_0\}, \tag{6.13}$$

where $\vec{e_0} = (1, \ldots, 1)^\tau$. *(This implies that* Ker $(\mathcal{L}) = \mathcal{K}_0$.*)*

(2) *The operator* $\mathcal{L}$ *is an invertible operator if restricted as follows:*

$$\mathcal{L} : \mathcal{K}_0^{\perp,1} \to \mathcal{K}_0^{\perp,2},$$

where

$$\mathcal{K}_0^{\perp,1} = \left\{ u \in (H^2(R))^N | \int_R uw^{'}(y)\vec{e_0} = 0 \right\},$$

$$\mathcal{K}_0^{\perp,2} = \left\{ u \in (L^2(R))^N | \int_R uw^{'}(y)\vec{e_0} = 0 \right\}.$$

The proof of Lemma 6.4 is technical and is delayed to Appendix C.

Now Theorem 6.1 (1) follows from Lemma 6.4 by the same proof as for Theorem 2.2 (3) of [58].

Item (2) of Theorem 6.1 follows the asymptotic analysis done at the beginning of this section.

To prove (3) of Theorem 6.1, we use the same argument as given in section 2 of [10], where the following eigenvalue problem was studied:

$$(6.14) \qquad \begin{cases} \epsilon^2 \Delta h - h + p u_\epsilon^{p-1} h - \frac{qr}{s+1+\tau\lambda_\epsilon} \frac{\int_\Omega u_\epsilon^{r-1} h}{\int_\Omega u_\epsilon^r} u_\epsilon^p = \lambda_\epsilon h \text{ in } \Omega, \\ h = 0 \text{ on } \partial\Omega, \end{cases}$$

where u_ϵ is a solution of the single equation

$$\begin{cases} \epsilon^2 \Delta u_\epsilon - u_\epsilon + u_\epsilon^p = 0 \text{ in } \Omega, \\ u_\epsilon > 0 \text{ in } \Omega, \; u_\epsilon = 0 \text{ on } \partial\Omega. \end{cases}$$

Here $1 < p < \frac{n+2}{n-2}$ if $n \geq 3$, and $1 < p < +\infty$ if $n = 1, 2$, $\frac{qr}{(s+1)(p-1)} > 1$, and $\Omega \subset R^n$ is a smooth bounded domain. If u_ϵ is a single interior peak solution, then it can be shown [56] that the limiting eigenvalue problem is an NLEP

$$(6.15) \qquad \Delta\phi - \phi + p w^{p-1}\phi - \frac{qr}{s+1+\tau\lambda_0} \frac{\int_{R^N} w^{r-1}\phi}{\int_{R^N} w^r} w^p = \lambda_0 \phi,$$

where w is the corresponding ground state solution in R^n:

$$\Delta w - w + w^p = 0, \quad w > 0 \text{ in } R^n, \quad w = w(|y|) \in H^1(R^n).$$

Dancer in [10] showed that if $\lambda_0 \neq 0$, $\text{Re}(\lambda_0) > 0$ is an unstable eigenvalue of (6.15), then there exists an eigenvalue λ_ϵ of (6.14) such that $\lambda_\epsilon \to \lambda_0$.

Now we follow his idea. Let $\lambda_0 \neq 0$ be an eigenvalue of (6.9) with $\text{Re}(\lambda_0) > 0$. First we note that, from the equation for ψ_ϵ, we can express ψ_ϵ in terms of $(\phi_{\epsilon,1}, \ldots, \phi_{\epsilon,N})$. Now we write the first equation for $(\phi_{\epsilon,1}, \phi_{\epsilon,2}, \ldots, \phi_{\epsilon,N})$ as follows:

$$(6.16) \quad \phi_{\epsilon,i} = -R_\epsilon(\lambda_\epsilon) \left[A M_\epsilon \sum_{j=1}^N k_{ij}(X_{\epsilon,j}\phi_{\epsilon,i} + \phi_{\epsilon,j} X_{\epsilon,i}) + A\psi_\epsilon \sum_{j=1}^N k_{ij} X_{\epsilon,i} X_{\epsilon,j} \right],$$
$$i = 1, \ldots, N,$$

where $R_\epsilon(\lambda)$ is the inverse of $-\Delta + (1 + \lambda_\epsilon)$ in $H^2(R)$ (which exists if $\text{Re}(\lambda_\epsilon) > -1$ or $\text{Im}(\lambda_\epsilon) \neq 0$). The important thing is that $R_\epsilon(\lambda_\epsilon)$ is a compact operator if ϵ is sufficiently small. The rest of the argument exactly follows that in [10]. For the sake of limited space, we omit the details here. □

7. Analysis of the vectorial NLEP and the proof of Theorem 2.2. In this section, we analyze the vectorial NLEP which we have obtained in (6.9):

$$\Delta\phi_i - \phi_i + \phi_i w + \sum_{j=1}^N b_{ij}\phi_j - \frac{\xi_i}{\sum_{i=1}^N \xi_i}\frac{2(1-\eta)}{1-\eta+\eta\beta_0}\frac{\sum_{j=1}^N \int_R \phi_j w}{\int_R w^2} w^2 \tag{7.1}$$
$$= \lambda_0\phi_i, \quad i = 1, \dots, N, \quad \phi_i \in H^2(R).$$

We will decouple it to a local eigenvalue problem with complex coefficients given in (5.1) and a scalar NLEP given in (5.2). Here assumptions (H2) and (H3) play a very important role. By Lemma 5.2, Lemma 5.3, and Theorem 6.1, we finish the proof of Theorem 2.2.

Proof of (1) *of Theorem* 2.2. Consider the case for $(X_\epsilon^s, M_\epsilon^s)$ and τ small. In this case, $0 \le \eta = \lim_{\epsilon\to 0} M_\epsilon(0) < \frac{1}{2}$. By Theorem 6.1 (1), if $\lambda_\epsilon = o(1)$, then $\lambda_\epsilon = 0$ and 0 is a simple eigenvalue. (The eigenspace is one-dimensional.) We need only to consider large eigenvalues. Let us assume that for a subsequence $\epsilon_n \to 0$ we have $\lambda_{\epsilon_n} \to \lambda_0$, where $\mathrm{Re}(\lambda_0) \ge 0$ and $\lambda_0 \ne 0$. We shall derive a contradiction.

By Theorem 6.1 (2), λ_0 is an eigenvalue of (7.1). First we take care of the nonlocal terms in (7.1). Adding the equations for $i = 1, \dots, N$ (using the assumption (H2)), we get

$$\Delta\left(\sum_{i=1}^N \phi_i\right) - \left(\sum_{i=1}^N \phi_i\right) + 2w\left(\sum_{i=1}^N \phi_i\right)$$
$$-\frac{2(1-\eta)}{\beta_0\eta + 1 - \eta}\frac{\int_R(\sum_{i=1}^N \phi_i)w}{\int_R w^2}w^2 = \lambda_0\sum_{i=1}^N \phi_i.$$

From Lemma 5.3 (1) we know that for $0 < \eta = \lim_{\epsilon\to 0} M_\epsilon(0) < \frac{1}{2}$ and τ small we have

$$\sum_{i=1}^N \phi_i = 0 \qquad \text{if } \mathrm{Re}\ (\lambda_0) \ge 0, \quad \lambda_0 \ne 0. \tag{7.2}$$

Therefore, the nonlocal terms in (7.1) all vanish, and (7.1) reduces to the following vectorial local eigenvalue problem:

$$\Delta\phi_i - \phi_i + w\phi_i + w\sum_{j=1}^N b_{ij}\phi_j = \lambda_0\phi_i, \quad \phi_i \in H^1(R), \quad i = 1, \dots, N. \tag{7.3}$$

To finish the proof, we have to transform this to Jordan form; we decompose

$$b_{ij} = \sum_{k,l=1}^N p_{ik}d_{kl}p_{lj}^{-1},$$

as in (2.17) of section 2, where d_{kl} has Jordan form.

Set

$$\Phi_i = \sum_{j=1}^N p_{ij}^{-1}\phi_j. \tag{7.4}$$

Then (7.3) can be expressed in terms of Φ as follows:

$$\Delta\Phi_i - \Phi_i + w\Phi_i + \sum_{j=1}^{N} d_{ij}\Phi_j w = \lambda_0\Phi_i, \quad i = 1, \dots, N. \tag{7.5}$$

We have to study the eigenvalue problems for each Jordan block separately.

Let σ be an eigenvalue of $\mathcal{B}$. By assumption (H3), $\sigma = 1$ is a simple eigenvalue of $\mathcal{B}$. Assume also that for those $\sigma \neq 1$ with $\sigma_R > 0$, it holds that $f(\sigma) < 0$.

For those eigenvalues $\sigma_k \neq 1, k > 1$, the corresponding ith component Φ_i of the eigenfunction satisfies

$$\Delta\Phi_i - \Phi_i + (1+\sigma_k)w\Phi_i = \lambda_0\Phi_i \tag{7.6}$$

with Re $(\lambda_0) \geq 0$.

By Lemma 5.2 (1), $\Phi_i = 0$ by our assumption on σ_k. Substituting this into the $(i-1)$th equation, we get (for the eigenfunction Φ_{i-1})

$$\Delta\Phi_{i-1} - \Phi_{i-1} + (1+\sigma_k)w\Phi_{i-1} = \lambda_0\Phi_{i-1}, \tag{7.7}$$

and by Lemma 5.2 (1) again we conclude that $\Phi_{i-1} = 0$. Continuing in this way, we see that those components of Φ corresponding to the Jordan block of σ_k all vanish.

Since $\sigma_1 = 1$ is a simple eigenvalue, we are left with the only possibility that $\Phi_1 \neq 0$. On the other hand, we have that

$$\sum_{j=1}^{N} \phi_j = \sum_{j=1}^{N} c_j\Phi_j, \tag{7.8}$$

where $c_j = \langle \vec{e}_0, \mathbf{p}_j \rangle$, where $\vec{e}_0 = (1, \dots, 1)^\tau$ and $\mathbf{p}_j$ is the jth column of $\mathcal{P}$. Note that $c_1 = \frac{\sum_{i=1}^{N} \xi_i}{\|\xi\|} > 0$.

Since $\sum_{j=1}^{N} \phi_j = 0$ and $\Phi_j = 0$, $j = 2, \dots, N$, we conclude from (7.8) that $\sum_{j=1}^{N} \phi_j = c_1\Phi_1 = 0$, and hence $\Phi_1 = \cdots = \Phi_N = 0$, which is a contradiction.

Therefore, Re $(\lambda_0) \geq 0$ is not possible. Thus we have Re $(\lambda_0) \leq -c_0 < 0$.

This proves (1) of Theorem 2.2. □

Proofs of (2) *and* (3) *of Theorem* 2.2. As before, we decompose $\mathcal{B} = \mathcal{P}\mathcal{D}\mathcal{P}^{-1}$ and let $\phi = \mathcal{P}\Phi$. The problem (7.1) is equivalent to the following:

$$\Delta\Phi - \Phi + w\Phi + w\mathcal{D}\Phi - \mathcal{P}^{-1}\xi \frac{2(1-\eta)}{(\beta_0\eta + 1 - \eta)(\sum_{i=1}^{N}\xi_i)} \frac{\sum_{i=1}^{N}\int_R w\phi_i}{\int_R w^2} w^2 = \lambda_0\Phi. \tag{7.9}$$

Note that

$$\mathcal{P}^{-1}\xi = \|\xi\|\vec{e}_1 \tag{7.10}$$

since ξ is the first eigenvector of $\mathcal{B}$, where $\vec{e}_1 = (1, 0, \dots, 0)^\tau$.

Therefore, (7.9) is decoupled into

$$\Delta\Phi_1 - \Phi_1 + 2w\Phi_1 - \frac{2\|\xi\|(1-\eta)}{(\beta_0\eta + 1 - \eta)(\sum_{i=1}^{N}\xi_i)} \frac{\sum_{i=1}^{N}\int_R w\phi_i}{\int_R w^2} w^2 = \lambda_0\Phi_1, \tag{7.11}$$

and

$$\Delta\Phi_i - \Phi_i + w\Phi_i + \sum_{j=1}^{N} d_{ij}\Phi_j w = \lambda_0\Phi_i, \quad i = 2, \ldots, K. \tag{7.12}$$

By (7.8) we have that

$$\int_R \sum_{i=1}^{N} w\phi_i = \sum_{j=1}^{N} c_j \int_R w\Phi_j. \tag{7.13}$$

We first prove (3) of Theorem 2.2. We consider $(X_\epsilon^l, M_\epsilon^l)$. In this case, $2(1-\eta) < 1$. By Lemma 5.3 (2), for any $\tau > 0$, there exist an eigenvalue $\lambda_0 > 0$ and an eigenfunction Φ_0 such that

$$\Delta\Phi_0 - \Phi_0 + 2w\Phi_0 - \frac{2(1-\eta)}{(\beta_0\eta + 1 - \eta)} \frac{\int_R w\Phi_0}{\int_R w^2} w^2 = \lambda_0\Phi_0, \quad \lambda_0 > 0.$$

Now we choose

$$\Phi_1 = \Phi_0, \quad \Phi_j = 0, \quad j = 2, \ldots, K. \tag{7.14}$$

Then $\mathbf{\Phi} = (\Phi_1, \ldots, \Phi_N)$ is a solution of (7.9) with $\lambda_0 > 0$. The corresponding $\phi = \mathcal{P}\Phi$ is a solution of (7.1) with $\lambda_0 > 0$. By Theorem 6.1 (3), we have the instability of $(X_\epsilon^l, M_\epsilon^l)$ for any $\tau > 0$.

This proves (3) of Theorem 2.2.

Finally, we prove (2) of Theorem 2.2. Consider $(X_\epsilon^s, M_\epsilon^s)$. Assume that there exist $\sigma_k \neq 1$ with $\mathrm{Re}(\sigma_k) > 0$ such that $f(\sigma_k) > 0$. By Lemma 5.2 (2), there exist an eigenvalue λ_0 with $\mathrm{Re}(\lambda_0) > 0$ and an eigenfunction Φ_0 such that

$$\Delta\Phi_0 - \Phi_0 + (1 + \sigma_k)w\Phi_0 = \lambda_0\Phi_0. \tag{7.15}$$

If σ_k is positive, we may choose λ_0 to be the principal eigenvalue given by Lemma 5.1 (3).

We choose $\Phi_k = \Phi_0$ and $\Phi_j = 0$ for $j \neq k$, $j \neq 1$. To choose Φ_1, we see that we have to solve (7.11), which becomes

$$\Delta\Phi_1 - \Phi_1 + 2w\Phi_1 - \frac{2(1-\eta)}{(\beta_0\eta + 1 - \eta)} \frac{\int_R w\Phi_1}{\int_R w^2} w^2 - \lambda_0\Phi_1 \tag{7.16}$$

$$= c_k \frac{2\|\xi\|(1-\eta)}{(\beta_0\eta + 1 - \eta)(\sum_{i=1}^{N} \xi_i)} \frac{\int_R w\Phi_0}{\int_R w^2} w^2.$$

To see that (7.16) is solvable, we note that (7.16) is equivalent to

$$\Delta\tilde{\Phi}_1 - \tilde{\Phi}_1 + 2w\tilde{\Phi}_1 - \lambda_0\tilde{\Phi}_1 = -\Lambda\lambda_0 w, \tag{7.17}$$

where

$$\tilde{\Phi}_1 = \Phi_1 - \Lambda w,$$

$$\Lambda = \frac{2(1-\eta)}{(\beta_0\eta+1-\eta)}\frac{\int_R w\Phi_1}{\int_R w^2} + c_k \frac{2\|\xi\|(1-\eta)}{(\beta_0\eta+1-\eta)(\sum_{i=1}^N \xi_i)}\frac{\int_R w\Phi_0}{\int_R w^2}.$$

If σ_k is not real, then $\mathrm{Im}(\lambda_0) \neq 0$, and so $L_0 - \lambda_0$ is invertible, where $L_0 = \Delta - 1 + 2w$. If σ_k is positive, then $\sigma_k \neq 1$ and $L_0 - \lambda_0$ is invertible. Thus (7.17) is solvable, and hence (7.16) is solvable. Going backward, we see that there exists a solution to (7.1) with $\mathbf{\Phi} = (\Phi_1, 0, \ldots, 0, \Phi_0, 0, \ldots, 0)$ and $\mathrm{Re}(\lambda_0) > 0$. Hence $(X_\epsilon^s, M_\epsilon^s)$ is unstable.

Item (2) of Theorem 2.2 is thus proved. □

Appendix A. Proof of Lemma 5.1. For (1), please see Lemma 4.1 of [51].

For (2), the fact that $\mu_1 = 1, \mu_2 = 2$ has already been proved in Lemma 4.1 of [51]. The exact value of μ_n can be computed using the same method as in the proof of Lemma 5.2. In fact, in this case, $\lambda = 0, \gamma = 1$, and hence the eigenvalues are given by

$$a = 2\gamma - \alpha = -(n-1),\, n = 1, 2, 3 \ldots,$$

where $\alpha^2 + \alpha - 6\mu = 0$. Thus $\mu_n = \frac{\alpha^2+\alpha}{6}, \alpha = n+1$.

Item (3) follows by the variational characterization of the eigenvalues:

$$-\lambda_1 = \inf_{\phi \in H^1(R), \phi \not\equiv 0} \frac{\int_R (\phi')^2 + \phi^2 - (1+\mu_R)w\phi^2}{\int_R \phi^2} < 0$$

since by the last inequality for $\phi = w$

$$-\lambda_1 \le -\mu_R \frac{\int_R w^3}{\int_R w^2} < 0.$$

This is the same analysis as in [61].

When $\mu_R = 1$, there exists only one positive eigenvalue (which is the principal one). See Lemma 1.2 of [56].

To prove (4), note that

$$\sigma = \sigma_R + i\sigma_I, \quad \phi = \phi_R + i\phi_I, \quad \lambda = \lambda_R + i\lambda_I,$$

and write the eigenvalue problem for real and imaginary parts separately:

$$\Delta\phi_R - \phi_R + (1+\sigma_R)w\phi_R - \sigma_I w\phi_I = \lambda_R\phi_R - \lambda_I\phi_I, \tag{A.1}$$

$$\Delta\phi_R - \phi_I + (1+\sigma_R)w\phi_I + \sigma_I w\phi_R = \lambda_R\phi_I + \lambda_I\phi_R. \tag{A.2}$$

Multiplying (A.1) by ϕ_R and (A.2) by ϕ_I, integrating over R, and adding, we get

$$\int_R [-({\phi_R}')^2 - \phi_R^2 + (1+\sigma_R)w\phi_R^2] + \int_R [-|{\phi_I}'|^2 - \phi_I^2 + (1+\sigma_R)w\phi_I^2]$$

$$= \lambda_R \int_R \phi_R^2 + \phi_I^2.$$

Since, in the last equation, the left-hand side is ≤ 0, we also get that the right-hand side is ≤ 0. Therefore, $\lambda_R \le 0$. Now assume that $\lambda_R = 0$. Then by (2) we get

$\phi_R = c_1 w$, $\phi_I = c_2 w$ (with $c_1, c_2 \in R$), and $\sigma_R = 0$. But this implies $\lambda_I = 0$, $\sigma_I = 0$, and we get $\lambda = 0$, contrary to what we assumed. Therefore, λ_R cannot be zero, and we conclude that Re $(\lambda) \leq -c_0 < 0$.

Appendix B. Proof of Lemma 5.4. In this appendix, we show how problem (5.2) can be reduced to (5.16).

Let ${}_AF_B$ be defined by (5.15). An important property of ${}_AF_B$ is the following integral property, whose proof can be found in [49]:

$$\text{(B.1)} \qquad {}_{A+1}F_{B+1}\left\{\begin{matrix} a_1, & a_2, & \ldots, & a_A, & c, & ; & \\ & & & & & & z \\ b_1, & b_2, & \ldots, & b_B, & d & ; & \end{matrix}\right\}$$

$$= \frac{\Gamma(d)}{\Gamma(c)\Gamma(d-c)} \int_0^1 t^{c-1}(1-t)^{d-c-1} {}_AF_B \left\{\begin{matrix} a_1, & a_2, & \ldots, & a_A & ; & \\ & & & & & tz \\ b_1, & b_2, & \ldots, & b_B & ; & \end{matrix}\right\} dt.$$

Let

$$f(\lambda) = \frac{2(1-\eta)}{\eta\sqrt{1+\tau\lambda}+1-\eta},$$

and let w be the unique solution of (2.1). Integrating (2.1), we have that

$$w^{'} = -\sqrt{w^2 - \frac{2}{3}w^3}.$$

First let us solve the following problem:

$$\text{(B.2)} \qquad \Delta\phi_0 - \phi_0 + 2w\phi_0 = w^2 + \lambda\phi_0, \quad \phi_0 \in H^2(R).$$

Since w is an even function, we may assume that ϕ_0 is also an even function. Let us denote the variable by t. Note that ϕ_0 is unique.

Set

$$\gamma = \sqrt{1+\lambda},$$

where we take the principal branch of $\sqrt{1+\lambda}$.

Then it is easy to see that problem (5.2) becomes

$$\text{(B.3)} \qquad \frac{1}{f(\lambda)} = \frac{\int_R w\phi_0}{\int_R w^2} = \frac{\int_0^{+\infty} w\phi_0 dt}{\int_0^{+\infty} w^2 dt}.$$

First let us set

$$\phi_0 = w^\gamma G.$$

Then, by some simple computations, G satisfies

$$\text{(B.4)} \qquad \frac{d^2G}{dt^2} + 2\gamma\frac{w^{'}}{w}\frac{dG}{dt} + \left(2 - \frac{\gamma}{3}(1+2\gamma)\right) wG = w^{1-\gamma}.$$

Next we perform the following change of variables:

$$z=\frac{2}{3}w. \tag{B.5}$$

Note that $w(0)=\frac{3}{2}$, and so z is a homeomorphism from $[0,+\infty]$ to $[0,1]$.

(We remark that here we take a different transformation as in [12]. Our transformation can be considered as a quadratic transformation for hypergeometric functions.)

By some lengthy computations, we obtain the following equation for $G(z)$:

$$z(1-z)G^{''}+(c-(a+b+1)z)G^{'}-abG=\left(\frac{3}{2}\right)^{2-\gamma}z^{1-\gamma}, \tag{B.6}$$

where

$$a=2+\gamma,\quad b=\gamma-\frac{3}{2},\quad c=1+2\gamma. \tag{B.7}$$

To solve (B.6), we take a power series

$$G(z)=z^s\sum_{k=0}^{+\infty}c_kz^k,$$

and, substituting it into (B.6), we obtain that

$$\sum_{k=0}^{+\infty}c_kz^{s+k-1}(s+k)(s+k-1+c)-\sum_{k=1}^{+\infty}c_kz^{s+k}(s+k+a)(s+k+b)=\left(\frac{3}{2}\right)^{2-\gamma}z^{1-\gamma}.$$

So

$$s-1=1-\gamma,\quad c_0s(s-1+c)=\left(\frac{3}{2}\right)^{2-\gamma},$$

$$c_k(s+k)(s+k-1+c)=c_{k-1}(s+k-1+a)(s+k-1+b).$$

By regrouping the coefficients, we have that

$$G(z)=\left(\frac{3}{2}\right)^{2-\gamma}(4-\gamma^2)^{-1}z^{2-\gamma}{}_3F_2\left\{\begin{matrix}1, & \frac{1}{2}, & 4 & ; & \\ & & & & z\\ & 3-\gamma, & 3+\gamma & ; & \end{matrix}\right\}. \tag{B.8}$$

Now we can compute

$$\begin{aligned}\int_0^{+\infty}w\phi_0dt&=\frac{3}{2}\int_0^1w^{1+\gamma}G(z)\frac{dz}{-w^{'}}\\&=\left(\frac{3}{2}\right)^{1+\gamma}\int_0^1z^{\gamma}(1-z)^{-\frac{1}{2}}G(z)dz\\&=\left(\frac{3}{2}\right)^{3}(4-\gamma^2)^{-1}\int_0^1z^2(1-z)^{-\frac{1}{2}}{}_3F_2\left\{\begin{matrix}1, & \frac{1}{2}, & 4 & ; & \\ & & & & z\\ & 3-\gamma, & 3+\gamma & ; & \end{matrix}\right\}dz.\end{aligned}$$

By (B.1), we obtain that

$$\int_0^{+\infty} w\phi_0 dt$$

$$= \left(\frac{3}{2}\right)^3 (4-\gamma^2)^{-1} \frac{\Gamma(3)\Gamma(\frac{1}{2})}{\Gamma(\frac{7}{2})} {}_4F_3 \left\{ \begin{matrix} 1, & \frac{1}{2}, & 4, & 3 & ; & \\ & & & & & 1 \\ & 3-\gamma, & 3+\gamma, & \frac{7}{2} & ; & \end{matrix} \right\}. \tag{B.9}$$

On the other hand, it is easy to compute that

$$\int_0^{+\infty} w^2 dt = \left(\frac{3}{2}\right)^2 \int_0^1 z^2(1-z)^{-\frac{1}{2}} dz = \left(\frac{3}{2}\right)^2 \frac{\Gamma(2)\Gamma(\frac{1}{2})}{\Gamma(2+\frac{1}{2})}. \tag{B.10}$$

By (B.9), (B.10), and (B.3), we obtain (5.16).

Appendix C. Proof of Lemma 6.4. We prove Lemma 6.4 in this appendix. We assume that the assumptions (H1)–(H3) are satisfied.

Proof of Lemma 6.4 (1). Recall that $L_0 = \Delta - 1 + 2w$. It is easy to check that $w^{'}\vec{e}_0 \in \text{Ker } (\mathcal{L})$. All we need to show is that the dimension of Ker $(\mathcal{L})$ is at most 1. To this end, let $\phi \in \text{Ker } (\mathcal{L})$. We first show that the nonlocal term vanishes. In fact, summing up all the equations and using the assumptions (H1) and (H2), we obtain

$$\Delta\left(\sum_{j=1}^N \phi_j\right) - \left(\sum_{j=1}^N \phi_j\right) + 2w\left(\sum_{j=1}^N \phi_j\right) - 2(1-\eta)\frac{\int_R w(\sum_{j=1}^N \phi_j)}{\int_R w^2} w^2 = 0$$

since $\beta_0 = \sqrt{1+\tau\lambda_0} = 1$.

That is,

$$\Delta\left(\sum_{j=1}^N \phi_j - cw\right) - \left(\sum_{j=1}^N \phi_j - cw\right) + 2w\left(\sum_{j=1}^N \phi_j - cw\right) = 0, \tag{C.1}$$

where

$$c = 2(1-\eta)\frac{\int_R w(\sum_{j=1}^N \phi_j)}{\int_R w^2}. \tag{C.2}$$

By Lemma 5.1 (1)

$$\sum_{j=1}^N \phi_j - cw \in \text{Ker } (L_0) = \text{span}\{w'\}.$$

So we have

$$\int_R w\left(\sum_{j=1}^N \phi_j - cw\right) = 0.$$

Substituting this relation into (C.2), we get

$$\int_R w\sum_{j=1}^N \phi_j = 0$$

since

$$2(1-\eta) \neq 1.$$

Thus in $\mathcal{L}$ the nonlocal term vanishes, and we obtain the following system of equations:

$$\Delta\phi_i - \phi_i + w\phi_i + \sum_{j=1}^{N} b_{ij} w\phi_j = 0, \quad i = 1, \dots, N.$$

We decompose

$$\mathbf{B} = \mathcal{P}\mathcal{D}\mathcal{P}^{-1}$$

as in (2.17) of section 2.

Set

$$\Phi_i = \sum_{j=1}^{N} p_{ij}^{-1} \phi_j.$$

Then the operator L can be expressed in terms of Φ as follows:

$$\Delta\Phi_i - \Phi_i + w\Phi_i + \sum_{j=1}^{N} d_{ij}\Phi_j w = 0.$$

If $1+\sigma \notin$ spec (EVP) (recall that (EVP) was defined in Lemma 5.1 (2)), then by the last line of the Jordan block corresponding to σ we get $\Phi_i = 0$ using Lemma 5.1. Using this in the previous line, we get $\Phi_{i-1} = 0$, etc. This implies that all components of Φ in the Jordan block corresponding to σ vanish.

If $1+\sigma \in$ spec (EVP), then by hypothesis (H3) we have $\sigma = 1$. By assumption (H3), the eigenvalue $\sigma = 1$ is simple. Since $\Phi_j = 0, j = 2, \dots, K$, we are left with Φ_1 only.

Now by Lemma 5.1 (1) we get $\Phi_1 \in \text{Ker } (L_0) = \text{span}\{w'\}$.

In conclusion, we have proved that except for $i = 1$, where $\Phi_1 = cw'$, $c \in R$, for all other $i = 2, \dots, N$, it holds that $\Phi_i = 0$. This implies that the dimension of Ker$\mathcal{L}$ is at most 1.

This finishes the proof of Lemma 6.4 (1). □

Proof of Lemma 6.4 (2). To show that $\mathcal{L}$ is invertible from $\mathcal{K}_0^{\perp,1} \to \mathcal{K}_0^{\perp,2}$, we need only to show that the conjugate operator of $\mathcal{L}$, denoted by $\mathcal{L}^*$, has the kernel $\mathcal{K}_0$. In fact, let $\phi \in \text{Ker}(\mathcal{L}^*)$. Then we have

$$\Delta\phi_i - \phi_i + \phi_i w + \sum_{j=1}^{N} b_{ji}\phi_j w$$

$$-2(1-\eta)\frac{\int_R w^2 \sum_{i=1}^{N} \xi_i\phi_i}{\int_R w^2 (\sum_{i=1}^{N} \xi_i)} w = 0, \quad i = 1, \dots, N.$$

Multiplying the ith equation by ξ_i and summing up all the equations, by (H1) we have the following:

$$\text{(C.3)}\ \Delta\left(\sum_{i=1}^{N} \xi_i\phi_i\right) - \left(\sum_{i=1}^{N} \xi_i\phi_i\right) + 2w\left(\sum_{i=1}^{N} \xi_i\phi_i\right) - 2(1-\eta)\frac{\int_R w^2 (\sum_{i=1}^{N} \xi_i\phi_i)}{\int_R w^2} w = 0.$$

Multiplying (C.3) by w and then integrating over R, we obtain

$$(1-2(1-\eta))\int_R w^2 \sum_{i=1}^N \xi_i\phi_i = 0.$$

Since $2(1-\eta) \neq 1$, it is easy to deduce that

$$\int_R w^2 \sum_{i=1}^N \xi_i\phi_i = 0.$$

This means that the nonlocal term vanishes. The rest of the proof of Lemma 6.4 (2) is similar to that of Lemma 6.4 (1), since spec $(\mathcal{B})$ = spec $(\mathcal{B}^\tau)$, and may be omitted. □

Acknowledgments. Matthias Winter thanks the Department of Mathematics at The Chinese University of Hong Kong for their kind hospitality. We thank two anonymous referees for their helpful comments which led us to improve this paper.

REFERENCES

[1] M. A. Andrade, J. C. Nuño, F. Morán, F. Montero, and G. J. Mpitsos, *Complex dynamics of a catalytic network having faulty replication into error-species*, Phys. D, 63 (1993), pp. 21–40.

[2] R. A. Barrio, C. Varea, J. L. Aragòn, and P. K. Maini, *A two-dimensional numerical study of spatial pattern formation in interacting Turing systems*, Bull. Math. Biol., 61 (1999), pp. 483–505.

[3] P. Bates and G. Fusco, *Equilibria with many nuclei for the Cahn-Hilliard equation*, J. Differential Equations, 160 (2000), pp. 283–356.

[4] P. Bates, E. N. Dancer, and J. Shi, *Multi-spike stationary solutions of the Cahn-Hilliard equation in higher-dimension and instability*, Adv. Differential Equations, 4 (1999), pp. 1–69.

[5] M. C. Boerlijst and P. Hogeweg, *Spiral wave structure in pre-biotic evolution: Hypercycles stable against parasites*, Phys. D, 48 (1991), pp. 17–28.

[6] K.-S. Cheng and W.-M. Ni, *On the structure of the conformal Gaussian curvature equation on R*, Duke Math. J., 62 (1991), pp. 721–737.

[7] P. Chacón and J. C. Nuño, *Spatial dynamics of a model for prebiotic evolution*, Phys. D, 81 (1995), pp. 398–410.

[8] M. B. Cronhjort and C. Blomberg, *Hypercycles versus parasites in a two-dimensional partial differential equations model*, J. Theoret. Biol., 169 (1994), pp. 31–49.

[9] M. B. Cronhjort and C. Blomberg, *Cluster compartmentalization may provide resistance to parasites for catalytic networks*, Phys. D, 101 (1997), pp. 289–298.

[10] E. N. Dancer, *On stability and Hopf bifurcations for chemotaxis systems*, Methods Appl. Anal., to appear.

[11] M. del Pino, P. Felmer, and J. Wei, *On the role of mean curvature in some singularly perturbed Neumann problems*, SIAM J. Math. Anal., 31 (1999), pp. 63–79.

[12] A. Doelman, A. Gardner, and T. J. Kaper, *Stability analysis of singular patterns in the 1-D Gray-Scott model: A matched asymptotic approach*, Phys. D, 122 (1998), pp. 1–36.

[13] D. S. Morgan, A. Doelman, and T. S. Kaper, *Stationary periodic patterns in the 1D Gray–Scott model*, Methods Appl. Anal., 7 (2000), pp. 105–149.

[14] A. Doelman, T. Kaper, and P. A. Zegeling, *Pattern formation in the one-dimensional Gray-Scott model*, Nonlinearity, 10 (1997), pp. 523–563.

[15] A. Doelman, R. A. Gardner, and T. J. Kaper, *Large stable pulse solutions in reaction-diffusion equations*, Indiana Univ. Math. J., 50 (2001), pp. 443–507.

[16] S.-I. Ei, Y. Nishiura, and B. Sandstede, in preparation.

[17] S.-I. Ei and Y. Nishiura, private communication.

[18] M. Eigen and P. Schuster, *The hypercycle. A principle of natural self organisation. Part A. Emergence of the hypercycle*, Naturwissenschaften, 64 (1977), pp. 541–565.

[19] M. Eigen and P. Schuster, *The hypercycle. A principle of natural self organisation. Part B. The abstract hypercycle*, Naturwissenschaften, 65 (1978), pp. 7–41.

[20] M. Eigen and P. Schuster, *The hypercycle. A principle of natural self organisation. Part C. The realistic hypercycle*, Naturwissenschaften, 65 (1978), pp. 341–369.

[21] M. Eigen and P. Schuster, *The Hypercycle: A Principle of Natural Self-Organisation*, Springer, Berlin, 1979.

[22] 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.

[23] P. Gray and S. K. Scott, *Autocatalytic reactions in the isothermal, continuous stirred tank reactor: Isolas and other forms of multistability*, Chem. Engrg. Sci., 38 (1983), pp. 29–43.

[24] P. Gray and S. K. Scott, *Autocatalytic reactions in the isothermal, continuous stirred tank reactor: Oscillations and instabilities in the system $A+2B \to 3B, B \to C$*, Chem. Engrg. Sci., 39 (1984), pp. 1087–1097.

[25] C. Gui, *Multi-peak solutions for a semilinear Neumann problem*, Duke Math. J., 84 (1996), pp. 739–769.

[26] 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.

[27] C. Gui and J. Wei, *Multiple interior peak solutions for some singularly perturbed Neumann problems*, J. Differential Equations, 158 (1999), pp. 1–27.

[28] C. Gui, J. Wei, and M. Winter, *Multiple boundary peak solutions for some singularly perturbed Neumann problems*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 249–289.

[29] J. K. Hale, L. A. Peletier, and W. C. Troy, *Exact homoclinic and heteroclinic solutions of the Gray–Scott model for autocatalysis*, SIAM J. Appl. Math., 61 (2000), pp. 102–130.

[30] J. K. Hale, L. A. Peletier, and W. C. Troy, *Stability and instability of Gray-Scott model: The case of equal diffusivities*, Appl. Math. Lett., 12 (1999), pp. 59–65.

[31] M. K. Kwong and L. Zhang, *Uniqueness of positive solutions of $\Delta u + f(u) = 0$ in an annulus*, Differential Integral Equations, 4 (1991), pp. 583–599.

[32] M. Kowalczyk, *Multiple spike layers in the shadow Gierer-Meinhardt system: Existence of equilibria and approximate invariant manifold*, Duke Math. J., 98 (1999), pp. 59–111.

[33] Y.-Y. Li, *On a singularly perturbed equation with Neumann boundary condition*, Comm. Partial Differential Equations, 23 (1998), pp. 487–545.

[34] K. Lindgren and M. G. Nordahl, *Evolutionary dynamics of spatial games*, Phys. D, 75 (1994), pp. 292–309.

[35] R. M. May, *Hypercycles spring to life*, Nature, 353 (1991), pp. 607–608.

[36] J. Maynard Smith, *Hypercycles and the origin of life*, Nature, 280 (1979), pp. 445–446.

[37] C. B. Muratov and V. V. Osipov, *Static spike autosolitons in the Gray-Scott model*, J. Phys. A, 33 (2000), pp. 8893–8916.

[38] 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.

[39] 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.

[40] W.-M. Ni and I. Takagi, *Point-condensation generated by a reaction-diffusion system in axially symmetric domains*, Japan J. Indust. Appl. Math., 12 (1995), pp. 327–365.

[41] W.-M. Ni, I. Takagi, and E. Yanagida, Tohoku Math. J. (2), to appear.

[42] Y. Nishiura and D. Ueyama, *A skeleton structure of self-replicating dynamics*, Phys. D, 130 (1999), pp. 73–104.

[43] Y. Nishiura and D. Ueyama, *Spatio-temporal chaos for the Gray-Scott model*, Phys. D, 150 (2001), pp. 137–162.

[44] W.-M. Ni and J. Wei, *On the location and profile of spike-layer solutions to singularly perturbed semilinear Dirichlet problems*, Comm. Pure Appl. Math., 48 (1995), pp. 731–768.

[45] W.-M. Ni, *Diffusion, cross-diffusion, and their spike-layer steady states*, Notices Amer. Math. Soc., 45 (1998), pp. 9–18.

[46] M. A. Nowak and R. M. May, *Evolutionary games and spatial chaos*, Nature, 359 (1993), pp. 826–829.

[47] J. Reynolds, J. Pearson, and S. Ponce-Dawson, *Dynamics of self-replicating spots in reaction-diffusion systems*, Phys. Rev. E (3), 56 (1997), pp. 185–198.

[48] J. Reynolds, J. Pearson, and S. Ponce-Dawson, *Dynamics of self-replicating patterns in reaction diffusion systems*, Phys. Rev. Lett., 72 (1994), pp. 2797–2800.

[49] L. J. Slater, *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge, UK, 1966.

[50] M. J. Ward, *An asymptotic analysis of localized solutions for some reaction-diffusion models in multi-dimensional domains*, Stud. Appl. Math., 97 (1996), pp. 103–126.

[51] J. Wei, *On the construction of single-peaked solutions to a singularly perturbed semilinear Dirichlet problem*, J. Differential Equations, 129 (1996), pp. 315–333.

[52] J. Wei, *On the boundary spike layer solutions of singularly perturbed semilinear Neumann problem*, J. Differential Equations, 134 (1997), pp. 104–133.

[53] J. Wei, *On the interior spike layer solutions for some singular perturbation problems*, Proc. Roy. Soc. Edinburgh Sect. A, 128 (1998), pp. 849–874.

[54] J. Wei, *On the interior spike layer solutions of singularly perturbed semilinear Neumann problem*, Tohoku Math. J., 50 (1998), pp. 159–178.

[55] J. Wei, *Uniqueness and critical spectrum estimates of boundary spike solutions*, Proc. Roy. Soc. Edinburgh Sect. A, to appear.

[56] J. Wei, *On single interior spike solutions of Gierer-Meinhardt system: Uniqueness, spectrum estimates and stability analysis*, European J. Appl. Math., 10 (1999), pp. 353–378.

[57] J. Wei, *Existence, stability and metastability of point condensation patterns generated by Gray-Scott system*, Nonlinearity, 12 (1999), pp. 593–616.

[58] J. Wei, *On two dimensional Gray-Scott model: Existence of single pulse solutions and their stability*, Phys. D, 148 (2001), pp. 20–48.

[59] J. Wei and M. Winter, *Stationary solutions for the Cahn-Hilliard equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 15 (1998), pp. 459–492.

[60] J. Wei and M. Winter, *Multiple boundary spike solutions for a wide class of singular perturbation problems*, J. London Math. Soc., 59 (1999), pp. 585–606.

[61] J. Wei and M. Winter, *On a two dimensional reaction-diffusion system with hypercyclical structure*, Nonlinearity, 13 (2000), pp. 2005–2032.

SIAM J. MATH. ANAL.
Vol. 33, No. 5, pp. 1090–1106

MEMORY DRIVEN INSTABILITY IN A DIFFUSION PROCESS*

BRIAN R. DUFFY[†], PEDRO FREITAS[‡], AND MICHAEL GRINFELD[†]

Abstract. We consider the n-dimensional version of a model proposed by Olmstead et al. [*SIAM J. Appl. Math.*, 46 (1986), pp. 171–188] for the flow of a non-Newtonian fluid in the presence of memory. We prove the existence of a global attractor and obtain conditions for the existence of a Lyapunov functional, which allows us to give a full description of this attractor in a certain region of the parameter space in the bistable case. We then study the stability and bifurcation of stationary solutions and, in particular, prove that for certain values of the parameters it is not possible to stabilize the flow by increasing a Rayleigh-type number. The existence of periodic and homoclinic orbits is also shown by studying the Bogdanov–Takens singularity obtained from a center manifold reduction.

Key words. parabolic systems, memory effects, non-Newtonian fluids

AMS subject classifications. 35B40, 35K55, 76E30

PII. S0036141001388592

1. Introduction.

In [14] Olmstead et al. studied the following equation as a model for the flow of a viscoelastic fluid:

$$u_t = \int_{-\infty}^{t} K(t,\ s) u_{xx}(x,s) ds + Rf(u), \tag{1.1}$$

on $x \in I = (0, \pi)$ and $t > 0$, subject to the Dirichlet boundary conditions

$$u(0,\ t) = u(\pi, t) = 0$$

and the initial condition $u(x, s) = u_0(x, s),\ s \leq 0$. Here $u(x, t)$ is the velocity of the fluid, R is a Rayleigh-type number, and f is taken to be cubic-like.

It is not hard to see (consulting [15], for example) that this equation cannot realistically describe the behavior of viscoelastic fluids. However, it is worth studying as it describes a reaction-diffusion process in a medium with memory (see [16] for a discussion of the diffusion processes in such media). In addition, it turns out to be an interesting mathematical object.

The bulk of [14] considers the case of the Jeffreys kernel

$$K(t,\ s) = \frac{1-\alpha}{\lambda} e^{-(t-s)/\lambda} + 2\alpha\delta(t-s),$$

with $\alpha \in [0,\ 1]$ and where δ denotes the Dirac delta function. Here λ has the physical meaning of relaxation time, and α is the ratio of retardation to relaxation times. If $\alpha = 0$, we have the Maxwell kernel, while if $\alpha = 1$ we recover the standard reaction-diffusion case. This case is also recovered in the limit $\lambda \to 0$, as is easily seen, for example, from (2.2).

*Received by the editors April 29, 2001; accepted for publication (in revised form) October 1, 2001; published electronically February 14, 2002.

http://www.siam.org/journals/sima/33-5/38859.html

[†]Department of Mathematics, University of Strathclyde, Livingstone Tower, 26 Richmond Street, Glasgow G1 1XH, UK (brd@maths.strath.ac.uk, michael@maths.strath.ac.uk).

[‡]Departamento de Matemática, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal (pfreitas@math.ist.utl.pt). This author's work was partially supported by FCT, Portugal, through project POCTI.

One of the characteristics of the Jeffreys kernel is that the stationary solutions depend only on the Rayleigh number. However, their stability does depend on the values taken by α and λ as well as R. This is, of course, due to the existence of other solutions, such as periodic solutions, which may bifurcate from the stationary solutions.

The stability and bifurcation analysis carried out in [14] was mainly based on the study of a set of ordinary differential equations obtained by a truncated Fourier series expansion. It is known that systems obtained by such procedures may display a behavior which is different from that of the original equation.

The purpose of the present paper is to carry out, in a rigorous way, the same type of study as in [14], but now for the n-dimensional version of (1.1), that is,

$$u_t = \int_{-\infty}^{t} K(t,s)\Delta u(x,s)ds + Rf(u), \quad x \in \Omega. \tag{1.2}$$

Here K is as before, Ω is a bounded domain in $\mathbb{R}^n$, and we consider Dirichlet boundary conditions. The function f is typically cubic-like, and whenever it will be necessary to be specific we will take f to be $u(1-u^2)$, although most of the results hold for more general classes of functions.

We begin by rewriting (1.2) as a strongly damped wave equation and then go on to show the existence of a Lyapunov functional for a certain region in parameter space. We then prove the existence of a global attractor with the help of invariant rectangles. This allows us to completely describe the attractor in the case where the Lyapunov functional exists, provided a bistable bifurcation diagram is assumed.

We then consider the stability and bifurcation of stationary solutions in terms of the parameters α, λ, and R. We relate the stability of the stationary solutions to that of the stationary solutions of a scalar parabolic equation and then go on to prove that there exist unbounded regions $\mathcal{Q}_1$ and $\mathcal{Q}_2$ in parameter space such that stable solutions of the parabolic equation remain stable in $\mathcal{Q}_1$, while in $\mathcal{Q}_2$ all stationary solutions of (1.1) become unstable. In particular, $\mathcal{Q}_2$ allows for R to be unbounded, which means that for certain values of α and λ it is not possible to stabilize the stationary solutions by increasing the value of the Rayleigh-type parameter R.

We remark that the results obtained in [14] for the finite-dimensional approximation actually state that for all positive α there exists a value of R, say R^*, such that for all $R > R^*$ there exists a nontrivial stable stationary solution. Thus, apart from the interest of this result as far as the behavior of solutions of (1.2) is concerned—large Rayleigh numbers do not necessarily stabilize the flow—it also illustrates the fact that finite-dimensional approximations, such as those considered in [14], may not represent the behavior of the infinite-dimensional system faithfully in this case. See section 7 for details.

The proof of these results relies on a careful analysis of the spectrum of the linearized operator around the stationary solutions and is based on the methods developed in [2, 3] and on the asymptotics of the stationary solution for large R. This enables us to reduce the problem to the study of a simpler self-adjoint operator, and we actually show that under certain conditions the linearization of (1.2) around a stationary solution always has positive *real* eigenvalues. This analysis also enables us to show that, by decreasing α and increasing λ, stationary solutions undergo a series of bifurcations due to pairs of complex eigenvalues crossing the imaginary axis from left to right, thus increasing the instability indices of these solutions; for a more precise statement, see Proposition 5.3.

Regarding the study of the bifurcations occurring in (1.2), we finally consider the linearized operator around the trivial solution for the values of the parameters for which there exists a double zero eigenvalue, and we make a center manifold reduction. This enables us to make a rigorous study of the behavior of solutions near that point. In particular, we show that this double zero eigenvalue corresponds to a Bogdanov–Takens singularity of the type studied in [7], and we obtain a complete description of the bifurcations occurring near this point. In the last section, the results obtained are then compared with those from [14].

2. The associated wave equation. As in [14], we begin by transforming the integro-differential equation (1.2) into a pair of coupled partial differential equations by introducing the auxiliary variable

$$v(x,t) = \frac{1-\alpha}{\lambda}\int_{-\infty}^{t} e^{-(t-s)/\lambda}u(x,s)\,ds. \tag{2.1}$$

This leads to the following system of equations:

$$\begin{cases} u_t = \alpha\Delta u + \Delta v + Rf(u), \\ \lambda v_t = (1-\alpha)u - v, \end{cases} \tag{2.2}$$

subject to homogeneous Dirichlet conditions in Ω and initial conditions at $t = 0$ derived from $u_0(x,s)$ in an obvious way. Note that, except in the case where α is one, we have from

$$v(x,0) = \frac{1-\alpha}{\lambda}\int_{-\infty}^{0} e^{s/\lambda}u(x,s)\,ds$$

that by taking

$$u(x,t) = e^{t/\lambda}\left[\left(1+\frac{2t}{\lambda}\right)u_0(x) - \frac{4t}{(1-\alpha)\lambda}v_0(x)\right], \quad t \leq 0,$$

for instance, it is possible to choose the initial conditions at $t = 0$ to be any given pair (u_0, v_0).

For the study below, it is convenient to rewrite (2.2) as a scalar equation

$$\lambda u_{tt} + \left[1-\lambda Rf'(u)\right]u_t = \Delta u + \alpha\lambda\Delta u_t + Rf(u), \tag{2.3}$$

together with Dirichlet boundary conditions and initial conditions of the form $u(x,0) = u_0(x)$ and $u_t(x,0) = u_1(x)$. In particular, keeping in mind that when α is one v vanishes identically, we now see that, as the parameter α goes from 0 to 1, we go from a damped wave equation to a reaction-diffusion equation, namely,

$$u_t = \Delta u + Rf(u). \tag{2.4}$$

It now becomes relatively simple to obtain conditions for the existence of a Lyapunov functional, and it will also be possible to apply the results from [3] directly—see sections 3 and 5, respectively.

This form of the equation also shows that nonzero values of α actually have the effect of adding a strong damping term to the wave equation corresponding to the Maxwell kernel ($\alpha = 0$).

Existence, uniqueness, and continuous dependence of solutions for (2.2) in the space $Z = \left[H^2(\Omega) \cap H_0^1(\Omega)\right] \times L^2(\Omega)$ now follow from standard results—see [12, 8], for instance.

3. Existence of a Lyapunov functional. Multiplying (2.3) by u_t and integrating over Ω gives

$$\begin{aligned}\frac{d}{dt}\int_\Omega \frac{\lambda}{2}u_t^2 dx &= -\int_\Omega \left[1-\lambda R f'(u)\right] u_t^2 dx + \int_\Omega u_t \Delta u dx \\ &\qquad + \alpha\lambda\int_\Omega u_t\Delta u_t dx + R\int_\Omega f(u)u_t dx \\ &= -\int_\Omega \left[1-\lambda R f'(u)\right] u_t^2 dx - \int_\Omega \nabla u\cdot\nabla u_t dx \\ &\qquad -\alpha\lambda\int_\Omega |\nabla u_t|^2 dx + R\frac{d}{dt}\int_\Omega F(u)dx,\end{aligned}$$

where F is a primitive of f and we have used integration by parts. By Poincaré's inequality, we thus have

$$\begin{aligned}\frac{d}{dt}\int_\Omega \frac{\lambda}{2}u_t^2 + \tfrac{1}{2}|\nabla u|^2 - RF(u)dx &= -\int_\Omega \left[1-\lambda R f'(u)\right] u_t^2 dx - \alpha\lambda\int_\Omega |\nabla u_t|^2 dx \\ &\le -\int_\Omega \left[1-\lambda R f'(u)+\alpha\lambda\sigma_p\right] u_t^2 dx,\end{aligned}$$

where σ_p is the principal eigenvalue of the Dirichlet Laplacian for the domain Ω.

Defining

$$M = \sup_{u\in\mathbb{R}}\left[f'(u)\right] \quad \text{and} \quad \Theta = \lambda(\alpha\sigma_p - MR)+1,$$

we have proven the following theorem.

THEOREM 3.1. *If Θ is positive, then the functional*

$$\mathcal{L}(u) = \int_\Omega \frac{\lambda}{2}u_t^2 + \frac{1}{2}|\nabla u|^2 - RF(u)dx$$

is strictly decreasing along trajectories of (2.3), *except at stationary solutions.*

If, for instance, F is bounded from above, then $\mathcal{L}$ is bounded from below, and, provided that all equilibria are isolated, we have that solutions converge to an equilibrium as t goes to infinity. In particular, this gives the following corollary.

COROLLARY 3.2. *Assume that $f(0)$ is zero, F is bounded from above, and both $\sigma_p - MR$ and Θ are positive. Then all solutions of* (2.3) *converge to zero.*

Proof. Since Θ is assumed to be positive and F is bounded from above, if we prove that there are no stationary solutions other than the trivial solution, we must have all solutions converging to zero.

The stationary solutions satisfy

$$\left\{\begin{array}{l}\Delta u + Rf(u) = 0,\\ v = (1-\alpha)u,\end{array}\right. \tag{3.1}$$

together with Dirichlet boundary conditions. If we multiply the first of these equations by Δu and integrate over Ω, we obtain

$$\int_\Omega (\Delta u)^2 dx = -R\int_\Omega f(u)\Delta u dx = R\int_\Omega f'(u)|\nabla u|^2 dx,$$

where in the integration by parts it was taken into account that $f(0)$ vanishes. We thus have that

$$\int_\Omega (\Delta u)^2 dx - MR\int_\Omega |\nabla u|^2 dx \leq 0.$$

On the other hand, using an argument similar to that used in [4] for the case of a compact manifold, we have that the functional

$$\int_\Omega (\Delta u)^2 dx - \xi\int_\Omega |\nabla u|^2 dx$$

is nonnegative for ξ less than or equal to σ_p. Hence, if $MR < \sigma_p$, it follows that u must vanish. ▯

4. Existence and structure of the attractor. In this section, we assume for simplicity that $f(u) = u(1-u^2)$, but clearly the results described are true for more general functions, as long as the necessary dissipativity conditions hold. We begin by showing the existence of invariant rectangles for α in $(0,1]$. This together with the results from [11] imply the existence of a compact connected attractor in $L^2 \times L^2$ for (2.2). For the parameter region where a Lyapunov functional exists, it is then possible to apply Mischaikow's results [13] to obtain the following theorem.

THEOREM 4.1. *The semigroup associated with* (2.3) *(and hence that associated with* (1.2)*) has a global attractor* $\mathcal{A}_{R,\lambda,\alpha}$ *in* L^2 *for* α *in* $(0,1]$. *Furthermore, if* Θ *is positive and the bifurcation diagram of stationary solutions is bistable, the flow on the attractor is semiconjugate to the flow on the Chaffee–Infante attractor* $\mathcal{A}_{R,\lambda,1}$.

Proof. We begin by rewriting system (2.2) so that the diffusion matrix is in diagonal form. For α in $(0,1]$, let

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & -\alpha \end{bmatrix} \begin{bmatrix} w \\ z \end{bmatrix}.$$

This yields

$$\begin{cases} w_t = \alpha\Delta w + h(w,z), \\ z_t = -\dfrac{1}{\alpha\lambda}\left[(1-\alpha)w + z\right], \end{cases}$$

where

$$h(w,z) = g(w+z) + \frac{1}{\lambda}z \quad\text{and}\quad g(t) = Rf(t) + \frac{1-\alpha}{\alpha\lambda}t.$$

Now consider the "kinetic equations" associated to this system, that is,

$$\begin{cases} w' = h(w,z), \\ z' = -\dfrac{1}{\alpha\lambda}\left[(1-\alpha)w + z\right]. \end{cases} \tag{4.1}$$

To prove the existence of invariant rectangles, we now proceed as in [18] and show that there are arbitrarily large rectangles in the (w,z) plane on the boundary of which the vector field of the "kinetic system" (4.1) points inward.

Let ℓ_1, ℓ_2, and ℓ_3 denote the lines in the plane wz defined by (see Figure 4.1)

$$z = (\alpha-1)w,\ z = -w - \sqrt{\frac{1}{3} + \frac{1}{3\alpha\lambda R}},\ \text{ and }\ z = -w + \sqrt{\frac{1}{3} + \frac{1}{3\alpha\lambda R}}.$$

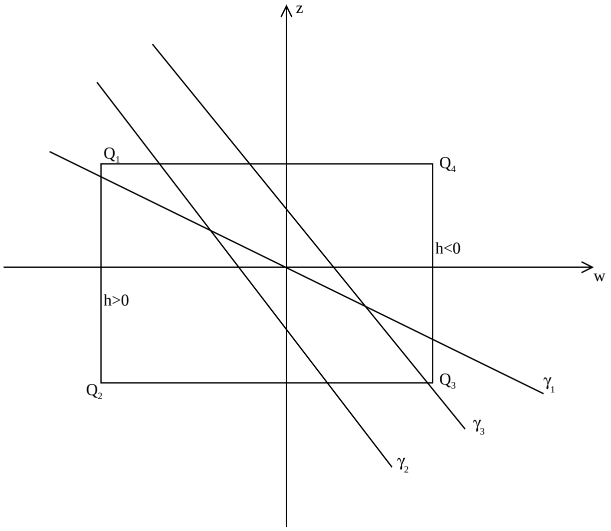


FIG. 4.1. *Invariant rectangles.*

First, note that for α in $(0, 1)$, γ_1 always intersects γ_2 and γ_3 and that z' is positive below γ_1 and negative above this line.

We have that

$$\frac{\partial h(w,z)}{\partial z} = g'(w+z) + \frac{1}{\lambda} = R(1-3(w+z)^2) + \frac{1}{\alpha\lambda}$$

is positive between γ_2 and γ_3 and negative otherwise. Furthermore, note that on γ_1 we have that $wh(w,z) = wh(w,(\alpha-1)w) = Rwf(\alpha w)$ is negative for $|w|$ large enough. It is thus possible to pick a point slightly above γ_1 (but still below γ_2) with negative w coordinate and such that h is positive at this point. In a similar way, it is possible to pick a point slightly below γ_1 (but above γ_3) with positive w coordinate and such that h is negative there. Denote these points by Q_1 and Q_3, respectively.

Now choose a point $Q_2 = -(w_1, z_1)$ in the third quadrant such that $-w_1$ is equal to the w coordinate of Q_1 and $-z_1$ is less than the z coordinate of Q_3. Then $h(Q_2)$ is positive on the segment Q_1Q_2 since $\partial h/\partial z$ is negative on that segment.

Now let Q_4 be a point in the first quadrant with z coordinate equal to that of Q_1 and w coordinate equal to that of Q_3. Then, since $h(Q_3)$ is negative and $\partial h/\partial z$ is negative above γ_3, it follows that h remains negative on the line segment Q_3Q_4.

In this way, we have that the rectangle $R = Q_1Q_2Q_3Q_4$ is such that $w' = h$ is positive on the left vertical side and negative on the right vertical side, and z' is positive on the lower horizontal side and negative on the upper horizontal side. Since the same construction starting from points Q_1 and Q_3 with more negative and more positive w coordinates, respectively, will still give a rectangle with these properties, we have shown the existence of arbitrarily large invariant rectangles for this equation. Verification of conditions (4.5)–(4.7) of [11] is now immediate, and we obtain the existence of a compact attractor in the space H of L^2 functions u and v taking almost all values in the invariant rectangle.

Hence, if the bifurcation diagram is bistable, as in the case considered in [14], under the conditions of Theorem 3.1 we can apply Mischaikow's results to conclude that the dynamics on the attractor of (2.3) (and thus of (1.2)) is semiconjugate to that of the Chaffee–Infante attractor corresponding to $\alpha = 1$. □

5. Stability of stationary solutions. In the specific case of a cubic-like function, such as that described in the introduction, the stationary problem (3.1) has been exhaustively studied in the literature—see, for instance, [10] for a survey of results regarding the existence of positive solutions for the elliptic problem associated with (2.4). In particular, we have that for $f(u) = u(1 - u^2)$ there is a branch of stable positive solutions bifurcating from the trivial solution and existing for all R larger than σ_p.

Linearizing (2.3) around a stationary solution corresponding to $(\bar{u}, (1-\alpha)\bar{u})$, we obtain the following eigenvalue problem in μ:

$$(1 + \alpha\lambda\mu)\Delta\phi + R(\lambda\mu + 1)a(x)\phi = \mu(\lambda\mu + 1)\phi \tag{5.1}$$

in Ω with Dirichlet boundary conditions and where $a(x) = f'(\bar{u})$. Methods that enable us to deal with problems of this type have been considered by the second author in [3]. From the results there, we immediately obtain the following proposition.

Proposition 5.1. *Let T be the operator obtained by linearizing* (2.2) *around a stationary solution. Then the spectrum of T consists of eigenvalues of finite multiplicity and exactly one point of the essential spectrum, $\mu = -1/(\alpha\lambda)$. Furthermore, there exist real numbers $\rho_0 > 0$ and $\pi/2 < \theta_0 < \pi$ such that the whole spectrum is contained inside a sector of the form $\{\lambda \in \mathbb{C} : \lambda = \rho_0 + \rho e^{i\theta}, |\theta| > \theta_0\}$.*

The idea behind the methods used in [3]—see also [2]—is to relate the eigenvalue problem (5.1) to a simpler self-adjoint problem. In order to do this, we rewrite (5.1) in the more familiar form

$$L_p\phi := \Delta\phi + pRa(x)\phi = \gamma\phi \tag{5.2}$$

by letting

$$\gamma = \frac{\mu(\lambda\mu + 1)}{\lambda\alpha\mu + 1} \quad \text{and} \quad p = \frac{\lambda\mu + 1}{\lambda\alpha\mu + 1}.$$

Consequently, we have that a real number $\mu \neq -1/(\alpha\lambda)$ will be a real eigenvalue of (5.1) provided that

$$\begin{cases} \mu = \dfrac{p - 1}{\lambda(1 - \alpha p)}, \\ \gamma = \dfrac{p(p - 1)}{\lambda(1 - \alpha p)}. \end{cases} \tag{5.3}$$

From the first of these equations we see that the values of the auxiliary parameter p that correspond to positive values of μ are $p \in (1, 1/\alpha)$. We thus look for intersections of the curve Γ defined by the second equation with the eigencurves of L_p, $\gamma_k(p)$, for $p \in (1, 1/\alpha)$. Note that, in particular, $\gamma_k(1)$ are the eigenvalues of the linearization of the reaction-diffusion equation (2.4).

A straightforward result is that for each positive eigenvalue $\gamma_j(1)$ of the operator L_1 there must exist at least one intersection between Γ and $\gamma_j(p)$ for p in $(1, 1/\alpha)$, and thus there corresponds at least a positive eigenvalue of problem (5.1). By noting that L_1 is precisely the operator associated with the linearization of (2.4) around the same stationary solution, we have the following theorem.

Theorem 5.2. *Let (u, v) be a stationary solution of* (2.2). *Then the dimension of the unstable manifold of (u, v) as a solution of this equation is greater than or equal to the dimension of the unstable manifold of u as a solution of* (2.4). *In particular, in the case where Ω is a ball, the only stationary solutions that may be stable are those which do not change sign.*

If Θ is positive, then the stability of the stationary solution is the same in both cases.

Proof. The first part is a simple consequence of the results in [3]. For the second part, multiply (5.1) by $\overline{\phi}$, and integrate by parts to obtain

$$-(1+\alpha\lambda\mu)\int_\Omega |\nabla\phi|^2 dx + R(\lambda\mu+1)\int_\Omega a(x)|\phi|^2 dx = \mu(\lambda\mu+1), \tag{5.4}$$

where it is assumed that ϕ has been normalized. Now separating μ into real and imaginary parts, we have from the equation for the imaginary parts that

$$-\alpha\lambda \mathrm{Im}(\mu)\int_\Omega |\nabla\phi|^2 dx + \lambda R \mathrm{Im}(\mu)\int_\Omega a(x)|\phi|^2 dx = \mathrm{Im}(\mu)\left[2\lambda\mathrm{Re}(\mu)+1\right].$$

It thus follows that either

$$-\alpha\lambda\int_\Omega |\nabla\phi|^2 dx + \lambda R\int_\Omega a(x)|\phi|^2 dx = 2\lambda\mathrm{Re}(\mu)+1$$

or $\mathrm{Im}(\mu)$ is zero. In the first case, we have that

$$2\lambda\mathrm{Re}(\mu) = -\alpha\lambda\int_\Omega |\nabla\phi|^2 dx + \lambda R\int_\Omega a(x)|\phi|^2 dx - 1 \le -\Theta,$$

and thus there are no nonreal eigenvalues with nonnegative real part when Θ is positive. This means that in this case the instability of a solution is determined only by the real eigenvalues.

Now assume that $\mathrm{Im}(\mu) = 0$. Then, for any positive eigenvalue of the linearization of the scalar parabolic equation (2.4) around the stationary solution, we have that the eigencurve γ of L_p passes through that point when p equals one. By continuity it must intersect the curve Γ defined by the second equation in (5.3) at some point $p \in (1, 1/\alpha)$. This shows that the same stationary solution is also unstable as a stationary solution of (2.3).

If, on the other hand, the stationary solution is asymptotically stable when considered as a stationary solution of (2.4), while it is unstable as a solution of (2.3), this means that there are at least two intersection points (counting multiplicities) of the eigencurve with the curve Γ from (5.3) for $p \in [1, 1/\alpha)$. There are now two cases to consider. First assume that the two intersections take place for p strictly larger than one. Since changing either of the two parameters α or λ does not affect the stationary solutions (and hence does not affect the eigencurve), it would now be possible, by making λ sufficiently small, or α sufficiently close to one, to change the curve Γ without changing the sign of Θ in such a way that no intersection takes place. This implies that the real eigenvalues have come together and become complex, which is not possible since we have seen that for positive values of Θ there are no nonreal eigenvalues with nonnegative real parts.

Now assume that one of the intersections takes place when p equals one. Then, again making λ sufficiently small or α sufficiently close to one, it is possible to make the positive eigenvalue cross zero and become negative, which would imply the existence of a bifurcation from a real eigenvalue giving rise to a stationary solution and contradicting the fact that the existence of stationary solutions does not depend on the parameters α or λ.

Finally, note that zero eigenvalues can occur only for intersections at p equal to one, and thus there exists a zero eigenvalue for (2.3) if and only if there exists one for (2.4). Also, if the stationary solution is stable for (2.4), then a zero eigenvalue must be simple as it is the first eigenvalue. Thus if there existed a zero eigenvalue with a higher multiplicity for (2.3), this would mean that a slight increase of the parameter λ would cause an eigenvalue to cross zero and become positive, while keeping Θ positive. This would again imply the existence of a bifurcation from a real eigenvalue, giving a contradiction. $\square$

From what has been seen, it seems likely that increasing λ and picking α close to zero will have a destabilizing effect on stationary solutions. This is indeed the case, as the following result shows.

Proposition 5.3. *Fix R, and let $(\bar{u}, (1-\alpha)\bar{u})$ be a stationary solution of* (2.2). *There exists a decreasing sequence of positive numbers α_k, $k = 1, \ldots,$ with $\alpha_1 < 1$, and an increasing sequence λ_k, $k = 1, \ldots,$ depending on α, such that for all $\alpha \in (\alpha_{k+1}, \alpha_k)$, if $\lambda > \lambda_k$, then there are at least $2k$ positive real eigenvalues of the linearized operator around the stationary solution $(\bar{u}, (1-\alpha)\bar{u})$.*

Proof. Denote by p_k^* the intersection point of the kth eigencurve for L_p with the positive part of the horizontal axis. From Proposition 2.3 in [2] we have that

$$\gamma_k(p) \geq \frac{\gamma_k(p_k^*) - \gamma_1(0)}{p_k^*} p + \gamma_1(0) = -\frac{\gamma_1(0)}{p_k^*} p + \gamma_1(0) \quad (p > p_k^*),$$

as $\gamma_k(p) > 0$ for all $p > p_k^*$. Thus, if $1/\alpha_k > p_k^*$, it follows that the eigencurve γ_k will have a portion above a straight line with positive inclination which is also above the p-axis on some subinterval of $(1, 1/\alpha_k)$. If we now take λ_k to be large enough, it follows that the graphs of γ_k and Γ must intersect at least twice for $p \in (1, 1/\alpha_k)$, giving rise to at least one pair of real positive eigenvalues. $\square$

Since, for α equal to one, the point spectrum coincides with that of the reaction-diffusion equation and the point of the essential spectrum is situated at $-1/\lambda$, we might expect that by making α close to one, solutions which are stable for the scalar equation (2.4) will remain stable.

Theorem 5.4. *Assume that $\bar{u}$ is a stable stationary solution of* (2.4), *and define*

$$\alpha_* = \frac{RM - 1/\lambda}{RM - \gamma_1(1)}.$$

If $\alpha > \alpha_*$, then $(\bar{u}, (1-\alpha)\bar{u})$ is a stable stationary solution of (2.3).

Proof. From (5.4) we have that μ satisfies the quadratic equation

$$\lambda\mu^2 + \left(1 + \alpha\lambda \int_\Omega |\nabla\phi|^2 dx - \lambda R \int_\Omega a(x)|\phi|^2 dx\right)\mu \\ + \int_\Omega |\nabla\phi|^2 dx - R\int_\Omega a(x)|\phi|^2 dx = 0,$$

and thus the stationary solution will be stable if and only if

$$\int_\Omega |\nabla\phi|^2 dx - R\int_\Omega a(x)|\phi|^2 dx > 0 \tag{5.5}$$

and

$$1 + \alpha\lambda \int_\Omega |\nabla\phi|^2 dx - \lambda R \int_\Omega a(x)|\phi|^2 dx > 0. \tag{5.6}$$

Since $\bar{u}$ is a stable solution of (2.4), we have that

$$\gamma_1(1) = \sup\left[-\int_\Omega |\nabla u|^2 dx + R\int_\Omega a(x)|u|^2 dx\right] < 0,$$

where the supremum is taken over all functions u in $H_0^1(\Omega)$ with $L^2(\Omega)$ norm 1. In particular,

$$\int_\Omega |\nabla\phi|^2 dx - R\int_\Omega a(x)|\phi|^2 dx \geq -\gamma_1(1) > 0,$$

showing that (5.5) is satisfied.

On the other hand, multiplying this last inequality by λ and adding 1 to both sides give

$$1 + \lambda\int_\Omega |\nabla\phi|^2 dx - \lambda R\int_\Omega a(x)|\phi|^2 dx \geq 1 - \gamma_1(1)\lambda,$$

and so

$$1 + \alpha\lambda\int_\Omega |\nabla\phi|^2 dx - \lambda R\int_\Omega a(x)|\phi|^2 dx \geq 1 - \gamma_1(1)\lambda + \lambda(\alpha - 1)\int_\Omega |\nabla\phi|^2 dx. \tag{5.7}$$

There are now two cases to consider. If

$$\int_\Omega |\nabla\phi|^2 dx < \frac{1 - \lambda\gamma_1(1)}{(1-\alpha)\lambda},$$

then the right-hand side of (5.7) is positive for all values of α, and we are done. Assume thus that

$$\int_\Omega |\nabla\phi|^2 dx \geq \frac{1 - \lambda\gamma_1(1)}{(1-\alpha)\lambda}.$$

Then

$$1 + \alpha\lambda\int_\Omega |\nabla\phi|^2 dx - \lambda R\int_\Omega a(x)|\phi|^2 dx \geq 1 + \frac{\alpha(1 - \lambda\gamma_1(1))}{1-\alpha} - \lambda RM,$$

and this last expression is positive for α in $(\alpha_*, 1)$. □

A straightforward consequence of this result is that for each stable solution of (2.4) there exists a value of α strictly smaller than 1, say α_{**}, such that the corresponding solution of (2.3) is stable for all α in $(\alpha_{**}, 1]$ and all λ. The following corollary gives precise details.

COROLLARY 5.5. *If $\bar{u}$ is as above and α is larger than $RM/(RM - \gamma_1(1))$, then $(\bar{u}, (1-\alpha)\bar{u})$ is stable for all positive values of λ.*

Note that both here and in Theorem 5.4 the term $\gamma_1(1)$ depends on the parameter R.

5.1. Positive solutions in balls. Theorem 5.2 states that only stationary solutions which are stable for the reaction-diffusion equation (2.4) may be stable for the wave equation (2.3). In this section, we show that there are regions in parameter space (α, λ) for which there are no stable solutions for any values of the parameter R. The proof uses estimates for the first eigencurve $\gamma_1(p)$ for large values of the parameter R.

In order to be able to obtain these estimates, we now restrict ourselves to the case where Ω is a ball and consider the specific case of $f(u) = u(1-u^2)$, although similar results could be obtained by the same methods for other types of functions, such as $f(u) = u(1-u^{p-1})$ for other values of p larger than one. Because Ω is a ball, only solutions which do not change sign can be stable when considered as stationary solutions of (2.4), and thus we now concentrate on branches of positive solutions. For this type of function, it is known that there exists a value of R, say R^*, such that for all R larger than R^* there exists one (and only one) positive stationary solution which is stable for (2.4).

THEOREM 5.6. *Assume that Ω is a ball. There exists α_0 in $(0,1)$, such that for all $0 \leq \alpha < \alpha_0$ there exists λ_0, depending on α, with the property that for all $\lambda > \lambda_0$ all stationary solutions of* (2.3) *are unstable for all values of R.*

By looking at the second equation in (5.3), we see that solutions (in p) corresponding to positive eigenvalues μ can exist only if the eigencurve $\gamma_k(p)$ is positive for some $p \in (1, 1/\alpha)$. As, in this case, all eigencurves are below the p-axis for $p \in (0,1)$, it becomes important to study the point $p_1^*(R)$ where the eigencurve $\gamma_1(p)$ intersects the real axis for positive p. Regarding this, we have the following lemma.

LEMMA 5.7. *Let Ω be the unit ball in $\mathbb{R}^n$. For each $k = 1, \ldots,$ there exists a unique positive value p_k^*, depending on R, such that $\gamma_k(p_k^*) = 0$ and $\gamma_k(p) > 0$ for all $p > p_k^*$. Furthermore, there exists an increasing sequence P_k, such that $p_k^*(R) < P_k$ for all $R > 1$. Furthermore, these values P_k can be chosen to be independent of $n \geq 1$.*

To establish the existence of, say, a value P_1 such that for all R the first eigencurve $\gamma_1(p)$ crosses the p-axis at values of $p < P_1$ requires two different pieces of information. The first ingredient is a uniformly valid matched asymptotics approximation of the (positive) solution u_R of the boundary value problem

$$\Delta u + Rf(u) = 0, \quad x \in \Omega, \ u = 0 \text{ on } \partial\Omega, \tag{5.8}$$

where $f(u) = u - u^3$. Note that by results of [10] such a positive solution exists for all $n \geq 1$ for sufficiently large R; furthermore, by the results of [5] this solution is radially symmetric.

The second ingredient of the proof is just the classical variational characterization of eigenvalues.

Matched asymptotics [9] give us the following representation:

$$u_R(r) \sim \tanh\left(\sqrt{\frac{R}{2}}(1-r)\right) + O(1/\sqrt{R}). \tag{5.9}$$

As $R \to \infty$, this solution is asymptotically equal to one everywhere in the ball far from the boundary, and it has a boundary layer close to $r = 1$ required to meet the boundary condition at $r = 1$. Note that the radially symmetric solution of (5.8) satisfies the boundary value problem

$$u_{rr} + \frac{n-1}{r}u_r + Rf(u) = 0, \quad u_r(0) = u(1) = 0, \tag{5.10}$$

and thus the term involving $(n-1)$ plays no role in the principal terms of either the outer or the inner expansions.

Now, by the variational characterization of eigenvalues, the eigenvalue $\gamma_1(p)$ of the problem

$$\Delta\phi + pRf'(u_R)\phi = \gamma\phi, \quad x \in \Omega,\ \phi = 0 \text{ on } \partial\Omega, \tag{5.11}$$

satisfies

$$\gamma_1(p) = \sup\left\{-\int_\Omega |\nabla\psi|^2 dx + pR\int_\Omega f'(u_R)\psi^2 dx\right\},$$

where the supremum is taken over all functions ψ in $H_0^1(\Omega)$ with $L^2(\Omega)$ norm 1. Hence if we show that for all arbitrarily large R there exist a function ψ_R and a value p^* independent of R such that

$$A = -\int_\Omega |\nabla\psi_R|^2 dx + p^*R\int_\Omega f'(u_R)\psi_R^2 dx > 0,$$

we have established the existence of P_1 and moreover $P_1 \le p^*$.

Consider the solution u_R with the asymptotic representation (5.9). Clearly, the function $f'(u_R)$ changes sign when $u_R = 1/\sqrt{3}$, that is, at the point

$$\overline{r} = 1 - \sqrt{\frac{2}{R}}\tanh^{-1}\left(1/\sqrt{3}\right) + O(1/R).$$

Consider the function $v_R(r)$ defined for $r \in [0,1]$ as follows:

$$v_R(r) = \begin{cases} \sin\frac{\pi(1-r)}{1-\overline{r}} & \text{if } r \in [\overline{r},\, 1], \\ 0 & \text{otherwise.} \end{cases}$$

Now let $\psi_R(r) = v_R(r)/\|v_R\|_2$.

Denoting by σ_n the surface area of the unit sphere, we have then

$$\frac{1}{\sigma_n}\|v_R\|^2 A = -\frac{\pi^2}{(1-\overline{r})^2}\int_{\overline{r}}^1 \cos^2\left[\frac{\pi(1-r)}{1-\overline{r}}\right] r^{n-1}dr$$
$$+pR\int_{\overline{r}}^1 \left[1 - 3\tanh^2\left(\sqrt{\frac{R}{2}}(1-r)\right) + O\left(\frac{1}{\sqrt{R}}\right)\right]\sin^2\left[\frac{\pi(1-r)}{1-\overline{r}}\right] r^{n-1}dr.$$

This can be rewritten as follows:

$$\frac{1}{\sigma_n}\|v_R\|^2 A = -\frac{\pi^2}{1-\overline{r}}\int_0^1 \cos^2(\pi y)(\overline{r} + y(1-\overline{r}))^{n-1}dy + pR(1-\overline{r})$$
$$\times\int_0^1 \left[1 - 3\tanh^2\left((1-y)\tanh^{-1}(1/\sqrt{3})\right) + O(1/\sqrt{R})\right](\overline{r} + y(1-\overline{r}))^{n-1}\sin^2(\pi y)dy.$$

Since $1-\overline{r} = O(1/\sqrt{R})$, this means that

$$\frac{1}{\sigma_n}\|v_R\|^2 A = \sqrt{R}\left[-\frac{\pi^2}{2\sqrt{2}\tanh^{-1}(1/\sqrt{3})} + p\sqrt{2}\tanh^{-1}(1/\sqrt{3})\beta + O(1/\sqrt{R})\right],$$

where

$$\beta = \int_0^1 \left[1 - 3\tanh^2\left((1-y)\tanh^{-1}(1/\sqrt{3})\right)\right] \sin^2(\pi y) dy \approx 0.3353.$$

Hence, if we choose $p > p^* = \pi^2/(4\beta(\tanh^{-1}(1/\sqrt{3}))^2) \approx 16.972$, we are guaranteed that, for all R sufficiently large, A is positive and so will be the eigenvalue $\gamma_1(p)$ for that value of p.

For P_k, $k > 1$, we use the same argument but employ functions

$$v_R^k(r) = \begin{cases} \sin\left[\dfrac{k\pi(1-r)}{1-\overline{r}}\right] & \text{if } r \in [\overline{r},\, 1], \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, v_R^k and v_R^l are orthogonal if $k \neq l$.

Remark 5.1. Numerically, the limiting value P_1 is approximately 10.8 in the one-dimensional case. We present a rough argument to approximate this value. By WKB reasoning we expect for large R the nonnegative eigenfunction to behave as our functions $\psi_R(v)$ above. In fact, we expect the sinusoidal part to be of the form $\sin\left[\sqrt{pR}(1-r)\right]$. On the other hand, we also expect from the variational formulation the sinusoidal part to be $\sin[\pi(1-r)/(1-\overline{r})]$. Thus in the limit we should have $\sqrt{P_1 R} = \pi/(1-\overline{r})$ as $R \to \infty$. But this implies that $P_1 = \pi^2/(2\tanh^{-1}(1/\sqrt{3})^2) \approx 11.38$.

Proof of Theorem 5.6. Let u_R be a stable stationary solution of the system (2.3). By Lemma 5.7 we have that, for any value of R on the interval $(R^*, +\infty)$, there exists a value P_1 such that the eigencurve $\gamma_1(p)$ intersects the horizontal axis at a value p_1^* smaller than P_1 and becomes positive for all values of p larger than p_1^*. This means that it is always possible to choose α_0^1 sufficiently small such that $1/\alpha$ is larger than p_1^*, independently of R, for α smaller than α_0^1. From the proof of Proposition 5.3, we know that the eigencurve is above the straight line joining $\gamma_1(0)$ and P_1. We may then pick λ_0^1 sufficiently large (and again independently of R) such that the curves γ_1 and Γ intersect at a point $(\overline{p}, \overline{\gamma}_1)$ with $\overline{p}$ in the interval $(1, 1/\alpha)$. This implies that the solution u_R is unstable.

Since the number of stable stationary solutions for the parabolic problem (2.4) with the nonlinearity f considered is finite for each value of R, we may now apply the above argument to all of these solutions and obtain values of α_0 and λ_0 for which all solutions are unstable when $\alpha \leq \alpha_0$ and $\lambda \geq \lambda_0$. □

6. The double zero eigenvalue. In this section, we consider a center manifold reduction for the double zero eigenvalue of the linearized equations around the trivial solution. We take $f(u) = u(1-u^2)$, but again it is clear that the argument holds for a much more general class of functions. In order to obtain a projection on the center manifold corresponding to a known form for the Bogdanov–Takens singularity, we rewrite (2.3) as

$$\begin{cases} u_t = w, \\ w_t = \dfrac{1}{\lambda}\Delta u + \alpha\Delta w + \dfrac{R}{\lambda} f(u) + \left[R f'(u) - \dfrac{1}{\lambda}\right] w. \end{cases}$$

This is equivalent to

$$\begin{bmatrix} u_t \\ w_t \end{bmatrix} = C \begin{bmatrix} u \\ w \end{bmatrix} + N(u,w), \tag{6.1}$$

where

$$C = \begin{bmatrix} 0 & 1 \\ \frac{1}{\lambda}(\Delta + R) & \alpha\Delta + \left(R - \frac{1}{\lambda}\right) \end{bmatrix} \quad \text{and} \quad N(u,w) = \begin{bmatrix} 0 \\ -\frac{R}{\lambda}u^3 - 3Ru^2w \end{bmatrix}.$$

The eigenvalue problem corresponding to the linearization around the trivial solution, which corresponds to the eigenvalues of the operator C, is given by

$$\begin{cases} \psi = \mu\phi, \\ \frac{1}{\lambda}\Delta\phi + \frac{R}{\lambda}\phi + \alpha\Delta\psi + \left(R - \frac{1}{\lambda}\right)\psi = \mu\psi. \end{cases}$$

In this case, it is possible to obtain an explicit expression for the eigenvalues, namely,

$$\mu_k^{\pm} = \frac{-\alpha\lambda\sigma_k + R\lambda - 1 \pm \sqrt{(1 - R\lambda + \alpha\lambda\sigma_k)^2 + 4(R - \sigma_k)\lambda}}{2\lambda}, \tag{6.2}$$

where, as before, σ_k denotes an eigenvalue of the Dirichlet Laplacian in Ω, that is,

$$\begin{cases} -\Delta v_k = \sigma_k v_k, & x \in \Omega, \\ v_k = 0, & x \in \partial\Omega. \end{cases}$$

As before, we shall denote the principal eigenvalue by σ_p and the corresponding normalized eigenfunction by v_p. It follows that there exists a double zero eigenvalue if and only if

$$\begin{cases} R = \sigma_k, \ k = 1, 2, \dots, \\ \alpha = 1 - \frac{1}{\lambda\sigma_k}. \end{cases}$$

Note that if α is 1, then it is not possible to satisfy the second equation, and thus there is no double zero eigenvalue in this case. On the other hand, for any α on $[0,1)$ it is possible to choose the remaining parameters R and λ so that zero is, in fact, a double eigenvalue.

Because of stability, we are interested in the case in which $k = 1$, which corresponds to

$$C_1 = \begin{bmatrix} 0 & 1 \\ \frac{1}{\lambda}(\Delta + \sigma_p) & \alpha\Delta + \left(\sigma_p - \frac{1}{\lambda}\right) \end{bmatrix}.$$

An eigenfunction for this operator and which corresponds to the zero eigenvalue is given by $\psi_1 = (v_p, 0)$. The other (generalized) eigenfunction is a solution of

$$C_1 \begin{bmatrix} u \\ w \end{bmatrix} = \begin{bmatrix} v_p \\ 0 \end{bmatrix}.$$

We thus obtain $\psi_2 = (0, v_p)$.

We now make the decomposition of the phase space Z into

$$X = \{rv_p, \ r = (r_1, r_2) \in \mathbb{R}^2\}$$

and $Y = X^{\perp}$. The linear space X is invariant by C, and, with respect to the canonical basis, it can be represented in X by

$$C = \begin{bmatrix} 0 & 1 \\ \epsilon_1 & \epsilon_2 \end{bmatrix},$$

where

$$\text{(6.3)}\qquad \begin{cases} \epsilon_1 = \dfrac{R-\sigma_p}{\lambda}, \\ \epsilon_2 = R - \dfrac{1}{\lambda} - \alpha\sigma_p. \end{cases}$$

In this case, M is equal to 1, and thus ϵ_2 is equal to $-\Theta/\lambda$. To see that it is possible to unfold the singularity by means of the parameters R and λ, for any value of α in $[0,1)$, we have to prove that the map taking (R,λ) to (ϵ_1,ϵ_2) is surjective near $(R^*,\lambda^*) = (\sigma_p, 1/(1-\alpha)\sigma_p)$. This follows from the fact that the Jacobian matrix for this map is given by

$$\begin{bmatrix} \dfrac{1}{\lambda} & \dfrac{\sigma_p - R}{\lambda^2} \\ 1 & \dfrac{1}{\lambda^2} \end{bmatrix},$$

and thus its determinant at (R^*,λ^*) is equal to $1/\lambda_*^3$, which is nonzero.

When $\epsilon \equiv (\epsilon_1,\, \epsilon_2) = 0$, the invariant subspace corresponding to the double eigenvalue is spanned by ψ_1 and ψ_2. In this case, $X = \operatorname{span}\{\psi_1,\psi_2\}$, and an element z of Z can be decomposed as $z = rv_p + y$, $r \in \mathbb{R}^2$, $y \in Y$. Following [1], we write (6.1) as

$$\begin{cases} \dot{r} = C(\epsilon)r + \overline{N}(rv_p + y), \\ \dot{y} = D(\epsilon)y + (I-\mathcal{P})N(rv_p+y), \\ \dot{\epsilon} = 0, \end{cases}$$

where $\mathcal{P}$ is the projection from Z on X, defined by

$$\mathcal{P}\begin{bmatrix} u \\ w \end{bmatrix} \;=\; \begin{bmatrix} \displaystyle\int_\Omega uv_p dx \\ \displaystyle\int_\Omega wv_p dx \end{bmatrix} v_p,$$

$D = (I-\mathcal{P})C$, and $\overline{N} : Z \to \mathbb{R}^2$ is given by

$$\overline{N}(u,w) \;=\; \langle \mathcal{P}N(u,w), v_p\rangle \;=\; \begin{bmatrix} 0 \\ -\dfrac{R}{\lambda}\displaystyle\int_\Omega (u^3 + 3\lambda u^2 w)v_p dx \end{bmatrix}.$$

The equation on the center manifold is given by

$$\dot{r} = C(\epsilon)r + \overline{N}(rv_p + h(r,\epsilon)),$$

where $h(r,\epsilon) = (h_1(r,\epsilon), h_2(r,\epsilon)) = O(r^2 + |\epsilon r|)$. The nonzero term in $\overline{N}$ can then be evaluated to give

$$-R\int_\Omega v_p^4 dx \left(\frac{1}{\lambda}r_1^3 + 3r_1^2 r_2\right) + h.o.t.$$

Hence

$$\begin{bmatrix} \dot{r_1} \\ \dot{r_2} \end{bmatrix} \;=\; \begin{bmatrix} 0 & 1 \\ \epsilon_1 & \epsilon_2 \end{bmatrix}\begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \;+\; \begin{bmatrix} 0 \\ -R\displaystyle\int_\Omega v_p^4 dx \left(\frac{1}{\lambda}r_1^3 + 3r_1^2 r_2\right) \end{bmatrix} + h.o.t.$$

Comparing this with the case in section 7.3 of [7], we see that this situation corresponds to having both parameters a_3 and b_3 in there equal to -1. The corresponding behavior is shown in Figures 7.3.7 and 7.3.9 in that section.

We have thus shown the following theorem.

THEOREM 6.1. *In a neighborhood of the double zero eigenvalue of the linearization of* (2.2) *around the trivial solution corresponding to* $(R^*, \lambda^*) = (\sigma_p, 1/((1-\alpha)\sigma_p))$, *the parameters R and λ unfold a Bogdanov–Takens singularity of the type*

$$\begin{cases} \dot{x} = y, \\ \dot{y} = \mu_1 x + \mu_2 y - x^3 - x^2 y \end{cases}$$

for (μ_1, μ_2) *in a neighborhood of* $(0, 0)$. *In particular, in a neighborhood of* (R^*, λ^*), *there are curves of homoclinic solutions of* (2.2) *and of blue sky bifurcations of periodic solutions.*

7. Conclusions. It remains to compare our results with those from [14] and the conclusions derived there.

First, note that the mechanism of restabilization of convection by increasing R, as seen in Figures 5 and 6 of [14, p. 185], cannot be correct, as the branches of stationary solutions that are stable for large R are spurious: these are branches of secondary bifurcations, which do not exist in the infinite-dimensional problem. Furthermore, Theorem 5.6 shows that there are values of λ and α for which there are no stable stationary solutions, independently of the value of R.

On the other hand, we have shown in Theorem 6.1 that the progression (as we increase R) of (a) Hopf bifurcation from the (stable) trivial (conduction) state followed by (b) a pitchfork bifurcation of unstable nontrivial (convection) states, (c) Hopf bifurcation from these nontrivial states (which leads the states without internal zeros to become stable), and (d) the blue sky annihilation of periodic solutions, which is clearly seen in Figures 5 and 6 of [14], persists in the infinite-dimensional system as well. This progression is consistent with the observations of [17] on the transition from conduction to steady convection in viscoelastic fluids via oscillatory convection.

The methods used in this paper were not able to determine whether or not another feature of the truncated dynamics of [14] persists in (2.2), namely, the loss of stability of a *stable* branch of equilibria due to an increase in R. This feature of the truncated dynamics leads to a conjecture that is stronger than the statement of Theorem 5.6, namely, that for any λ and $\alpha < 1$, increasing R will destabilize convection. Clearly, increasing R would eventually make Θ negative. Such a conjecture is also consistent with (again, finite-dimensional) numerics of [6].

Finally, we point out that the case corresponding to the Maxwell kernel ($\alpha = 0$) poses some interesting open problems. In particular, we see from (6.2) that along the line defined by $\lambda R = 1$ there is an infinite-dimensional center manifold.

REFERENCES

[1] J. Carr, *Applications of Centre Manifold Theory*, Springer-Verlag, New York, 1981.

[2] P. Freitas, *On some eigenvalue problems related to the wave equation with indefinite damping*, J. Differential Equations, 127 (1996), pp. 320–335.

[3] P. Freitas, *Eigenvalue problems for the wave equation with strong damping*, Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 755–771.

[4] P. Freitas, *On minimal eigenvalues of Schrödinger operators on compact manifolds*, Comm. Math. Phys., 217 (2001), pp. 375–382.

[5] B. Gidas, W.-M. Ni, and L. Nirenberg, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209–243.

[6] M. Grinfeld, *Dynamics of a model equation in viscoelasticity*, in Proceedings EQUADIFF '91, vol. II, World Scientific, Singapore, 1993, pp. 568–572.

[7] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurcation of Vector Fields*, Springer-Verlag, New York, 1983.

[8] J. Hale, *Asymptotic Behavior of Dissipative Systems*, Math. Surveys Monogr. 25, AMS, Providence, RI, 1988.

[9] M. H. Holmes, *Introduction to Perturbation Methods*, Springer-Verlag, New York, 1998.

[10] P.-L. Lions, *On the existence of positive solutions of semilinear elliptic equations*, SIAM Rev., 24 (1982), pp. 441–467.

[11] M. Marion, *Finite-dimensional attractors associated with partly dissipative reaction-diffusion systems*, SIAM J. Math. Anal., 20 (1989), pp. 816–844.

[12] P. Massatt, *Limiting behavior for strongly damped nonlinear wave equations*, J. Differential Equations, 48 (1983), pp. 334–349.

[13] K. Mischaikow, *Global asymptotic dynamics of gradient-like bistable equations*, SIAM J. Math. Anal., 26 (1995), pp. 1199–1224.

[14] W. E. Olmstead, S. H. Davis, S. Rosenblat, and W. L. Kath, *Bifurcation with memory*, SIAM J. Appl. Math., 46 (1986), pp. 171–188.

[15] M. Renardy, *Mathematical analysis of viscoelastic flow*, in Annual Review of Fluid Mechanics, Annu. Rev. Fluid Mech. 21, Annual Reviews, Palo Alto, CA, 1989, pp. 21–36.

[16] R. E. Showalter and N. J. Walkington, *A hyperbolic Stefan problem*, Quart. Appl. Math., 45 (1987), pp. 769–781.

[17] M. Sokolov and R. I. Tanner, *Convective stability of a general viscoelastic fluid heated from below*, Phys. Fluids, 15 (1972), pp. 534–539.

[18] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, Berlin, 1983.

SIAM J. Math. Anal.
Vol. 33, No. 5, pp. 1107–1116

QUADRILATERAL MACROELEMENTS*

MING-JUN LAI† AND LARRY L. SCHUMAKER‡

Abstract. Macroelements of smoothness C^r depending only on natural derivative information are constructed on triangulated quadrilaterals for all $r \geq 1$.

Key words. bivariate splines, approximation order, quadrangulation, scattered data interpolation

AMS subject classifications. 41A15, 41A63, 41A25, 65D10

PII. S0036141000377134

1. Introduction. Suppose that Q is a convex quadrilateral and that $\triangle_Q$ is the triangulation obtained by splitting Q into four triangles by drawing in the two diagonals. Let v_Q be the intersection of the diagonals.

The first macroelement on $\triangle_Q$ was the C^1 piecewise cubic macroelement constructed in [8, 17]; see also [6]. Later, a class of C^r macroelements on $\triangle_Q$ was constructed in [14, 15]; see also [5]. The aim of this paper is to improve these higher smoothness macroelements by removing unnatural degrees of freedom.

The macroelements in [15] are based on the superspline spaces

$$\begin{aligned} &\mathcal{S}_{6m+1}^{2m+1,3m}(\triangle_Q), \qquad r = 2m, \\ &\mathcal{S}_{6m+3}^{2m+1,3m+1}(\triangle_Q), \qquad r = 2m+1, \end{aligned} \tag{1.1}$$

where, in general, if $\triangle$ is a triangulation of a domain Ω,

$$\mathcal{S}_d^{r,\rho}(\triangle) := \{s \in C^r(\Omega) \ : \ s \text{ is a piecewise polynomial of degree } d \text{ on } \triangle, \ s \in C^\rho(v) \text{ for all vertices } v\}. \tag{1.2}$$

As usual, $C^\rho(v)$ means that all polynomials on triangles sharing the vertex v have common derivatives up to order ρ at that vertex.

In this paper, we will make use of certain subspaces of the superspline spaces (1.1) which satisfy additional supersmoothness at the vertex v_Q as well as some other special smoothness conditions.

The paper is organized as follows. In section 2, we review some well-known Bernstein–Bézier notation and state a key lemma. In section 3, we discuss the case where r is even, and in section 4, we illustrate it with several examples. The case where r is odd is treated in section 5 and illustrated in section 6. Section 7 contains results on the corresponding global spline spaces, including their approximation power. Concluding remarks can be found in section 8.

*Received by the editors August 24, 2000; accepted for publication (in revised form) August 24, 2001; published electronically February 14, 2002.

http://www.siam.org/journals/sima/33-5/37713.html

†Department of Mathematics, University of Georgia, Athens, GA 30602 (mjlai@math.uga.edu). This author's work was supported by the National Science Foundation under grant DMS-9870178.

‡Department of Mathematics, Vanderbilt University, Nashville, TN 37240 (s@mars.cas.vanderbilt.edu). This author's work was supported by the National Science Foundation under grant DMS-9803340 and by the U.S. Army Research Office under grant DAAD-19-99-1-0160.

2. Notation and preliminaries. We make use of standard Bernstein–Bézier techniques. Given a triangle $T := \langle u_1, u_2, u_3 \rangle$ and an integer d, let

$$\xi^T_{ijk} := \frac{(iu_1 + ju_2 + ku_3)}{d}, \qquad i + j + k = d,$$

be the corresponding *domain points*. We will work with the usual rings and disks of domain points defined by

$$R^T_n(u_1) := \{\xi^T_{ijk} \ : \ i = d - n\},$$
$$D^T_n(u_1) := \{\xi^T_{ijk} \ : \ i \geq d - n\},$$

with similar definitions at the other vertices of T. It is well known (see [10] for explicit formulae) that specifying the B-coefficients in the disk $D^T_n(u_1)$ of a polynomial p is equivalent to specifying the derivatives $D^\nu_x D^\mu_y p(u_1)$ for $0 \leq \nu + \mu \leq n$.

Given a triangulation $\triangle$, we are interested in spline spaces which are subsets of the space $\mathcal{S}^0_d(\triangle)$ of splines which are globally C^0 and are piecewise polynomials of degree d. The corresponding set of domain points is defined to be the union of the $\{\xi^T_{ijk}\}$ as T runs over the triangles of $\triangle$, where points on edges are not repeated. We recall that a *minimal determining set (MDS)* for a spline space $\mathcal{S} \subseteq \mathcal{S}^0_d(\triangle)$ is a subset $\mathcal{M}$ of the domain points associated with $\mathcal{S}^0_d(\triangle)$ such that every spline $s \in \mathcal{S}$ is uniquely determined by the set of B-coefficients which are identified with the points of $\mathcal{M}$.

We shall make extensive use of certain linear functionals defined by smoothness conditions between polynomials of degree d on adjoining triangles. These were introduced in [2], but we repeat their definition here for convenience. Suppose that $T := \langle u_1, u_2, u_3 \rangle$ and $\tilde{T} := \langle u_4, u_3, u_2 \rangle$ are two adjoining triangles which share the edge $e := \langle u_2, u_3 \rangle$. Let s be a function whose restrictions to T and $\tilde{T}$ are polynomials of degree d. Let c_{ijk} and $\tilde{c}_{ijk}$ be the coefficients of the B-representations of s_T and $s_{\tilde{T}}$, respectively. Then, for any $n \leq m \leq d$, we define

$$\tau^n_{m,e} s := \tilde{c}_{n,m-n,d-m} - \sum_{i+j+k=n} c_{i,j+d-m,k+m-n} B^n_{ijk}(u_4), \tag{2.1}$$

where B^n_{ijk} are the Bernstein polynomials of degree n on the triangle T.

The following lemma [2] can be used to compute certain coefficients of s on the ring $R_m(u_2)$ assuming that an appropriate set of smoothness conditions across the edge e is satisfied.

LEMMA 2.1. *Suppose s is a piecewise polynomial of degree d defined on $T \cup \tilde{T}$ and that $d, m, p, q, \tilde{q}$ are integers with $0 \leq q, \tilde{q}$, $-1 \leq p \leq q, \tilde{q}$, and $q + \tilde{q} - p \leq m \leq d$. Suppose that*

$$\tau^n_{m,e} s = 0, \qquad p + 1 \leq n \leq q + \tilde{q} - p, \tag{2.2}$$

and that all of the coefficients c_{ijk} involved in these smoothness conditions are known except for

$$\begin{aligned} c_\nu &:= c_{\nu,d-r,r-\nu}, \qquad && \nu = p+1, \ldots, q, \\ \tilde{c}_\nu &:= \tilde{c}_{\nu,r-\nu,d-r}, \qquad && \nu = p+1, \ldots, \tilde{q}. \end{aligned} \tag{2.3}$$

Then the coefficients (2.3) *are uniquely determined by* (2.2).

3. The case $r = 2m$. Let Q be a quadrilateral with vertices $v_1, \dots, v_4$ in counterclockwise order. We define the triangles $T^{[i]} := \langle v_Q, v_i, v_{i+1}\rangle$ and edges $e_i := \langle v_i, v_Q\rangle$ for $i = 1, 2, 3, 4$, where $v_5 = v_1$ and v_Q is the point where the two diagonals of Q intersect.

THEOREM 3.1. *Given $r = 2m$, let $\mathcal{S}_r(\triangle_Q)$ be the linear subspace of all splines s in $\mathcal{S}_{6m+1}^{2m+1,3m}(\triangle_Q)$ that satisfy the following set of additional smoothness conditions:*

$$s \in C^{4m}(v_Q), \tag{3.1}$$

$$\tau_{3m+i+1,e_l}^{2m+1+j} s = 0, \qquad 1 \le j \le 2i, \quad 1 \le i \le m-1, \quad l = 1, 2, 3, 4, \tag{3.2}$$

and

$$\tau_{4m+1,e_1}^{2m+1+j} s = 0, \qquad 1 \le j \le 2m. \tag{3.3}$$

Then

$$\dim \mathcal{S}_r(\triangle_Q) = 26m^2 + 22m + 4.$$

Moreover, the following set $\mathcal{M}_r$ of domain points is an MDS:

(1) $D_{3m}^{T^{[i]}}(v_i)$ *for $i = 1, 2, 3, 4$,*

(2) $\{\xi_{j,3m,3m-j+1}^{T^{[i]}}, \dots, \xi_{j,3m-j+1,3m}^{T^{[i]}}\}$ *for $j = 1, \dots, 2m$ and $i = 1, 2, 3, 4$.*

Proof. First, we show that $\mathcal{M}_r$ is a determining set. Suppose that $s \in \mathcal{S}_r(\triangle_Q)$ and that we have set the coefficients of s corresponding to all domain points in $\mathcal{M}_r$. Then, using the usual smoothness conditions, we solve for the unset coefficients corresponding to domain points in the disks $D_{3m}(v_i)$ for $i = 1, 2, 3, 4$.

We now make use of Lemma 2.1. First, we compute the coefficients on the rings $R_{3m+i+1}(v_l)$ for $i = 0, \dots, m-1$ and $l = 1, 2, 3, 4$. On the ring $R_{3m+i+1}(v_l)$, this involves solving a system of $2(m+i)+1$ linear equations. Note that the spline satisfies all of the smoothness conditions required for the lemma since they either are already implicit in the supersmoothness of the space or have been explicitly enforced in the definition of $\mathcal{S}_r(\triangle_Q)$.

Using the lemma, we now compute the coefficients on $R_{4m+1}(v_1)$. We now start a sequence of calculations. First, we compute the $4m$ unset coefficients corresponding to points on the edge E_0, where, in general, E_i is the set of domain points in $T^{[1]} \cup T^{[2]}$ at a distance i from the edge $\langle v_1, v_3\rangle$. Then we compute the $4m$ coefficients corresponding to points in the set $\tilde{E}_0$, where $\tilde{E}_i$ is the set of domain points in $T^{[2]} \cup T^{[3]}$ at a distance i from the edge $\langle v_2, v_4\rangle$. The remaining coefficients in $T^{[1]} \cup T^{[2]} \cup T^{[3]}$ are computed by alternately working on the sets E_i and $\tilde{E}_i$ for $i = 1, \dots, r$. Finally, we compute the remaining coefficients in $T^{[4]}$ from the C^r smoothness conditions.

We have shown that all coefficients of s are determined by those corresponding to the domain points in the set $\mathcal{M}_r$. This shows that $\mathcal{M}_r$ is a determining set.

To see that $\mathcal{M}_r$ is an MDS, we consider $\mathcal{S}_{6m+1}^{2m+1}(\triangle_Q) \cap C^{4m}(v_Q)$. By Theorem 2.2 in [18] the dimension of this space is $32m^2+18m+4$. Our space $\mathcal{S}_r(\triangle_Q)$ is the subspace which satisfies the $4m^2 - 2m$ special conditions (3.2)–(3.3) and the supersmoothness $C^{3m}(v_i)$ for $i = 1, 2, 3, 4$. Enforcing the supersmoothness requires an additional $2m^2 - 2m$ conditions. Thus

$$\begin{aligned}(32m^2 + 18m + 4) - (4m^2 - 2m) - (2m^2 - 2m)& \\ \le \dim \mathcal{S}_r(\triangle_Q) \le \#\mathcal{M}_r = 26m^2 + 22m + 4.&\end{aligned}$$

Since the expression on the left equals the one on the right, we conclude that it is equal to the dimension of $\mathcal{S}_r(\triangle_Q)$, and $\mathcal{M}_r$ is an MDS. □

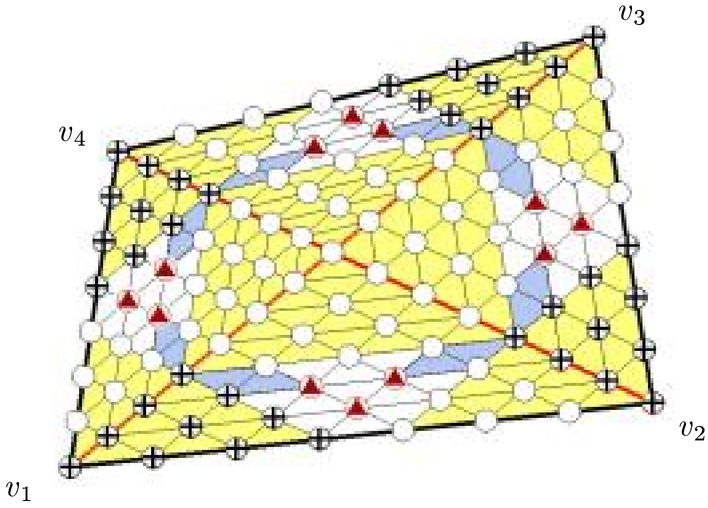


Fig. 1. *The C^2 macroelement.*

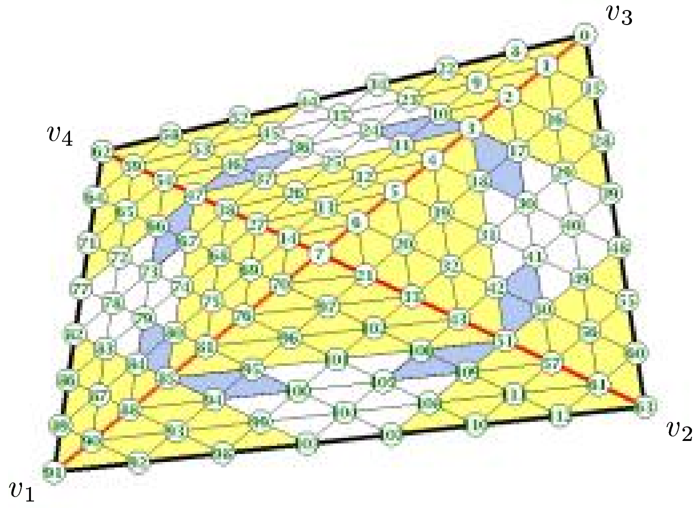


Fig. 2. *Domain points for the C^2 macroelement.*

4. Examples. In this section, we illustrate the construction of section 3.

Example 4.1. *The space $\mathcal{S}_2(\triangle_Q)$ is the subspace of $\mathcal{S}_7^3(\triangle_Q) \cap C^4(v_Q)$ that satisfies the two special smoothness conditions corresponding to τ_{5,e_1}^4 and τ_{5,e_1}^5.*

Discussion. The dimension of $\mathcal{S}_2(\triangle_Q)$ is 52, and the MDS for this macroelement is shown in Figure 1. It consists of 10 points in each of the disks $D_3(v_i)$ (marked with crosses) and 3 points corresponding to item (2) of Theorem 3.1 for each edge of Q (marked with triangles). After setting the coefficients in the MDS, the remaining coefficients are computed in the following order. First, we use C^3 smoothness to compute

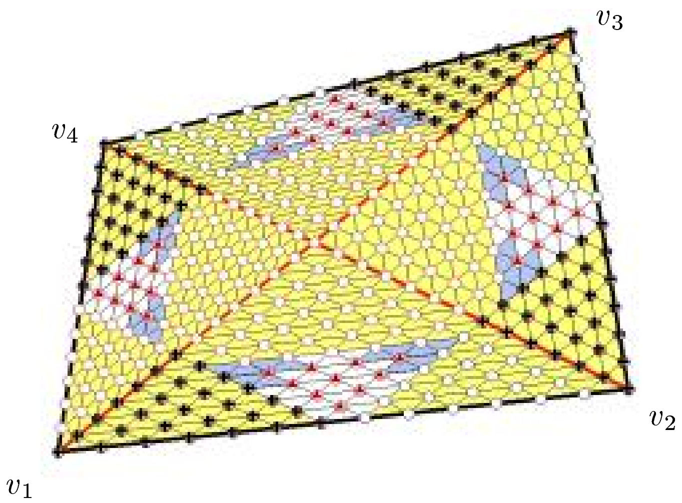


Fig. 3. *The C^4 macroelement.*

the coefficients numbered 80,81,95 in Figure 2, followed by those numbered 106,43,42, then 18,4,11, and 37,38,67. Using the two special smoothness conditions, we can now compute the coefficients numbered 74,75,76,96,101. Then, using C^4 smoothness, we compute the coefficients numbered 70,7,6,5 (lying in the set E_0) and 33,21,14,27 (lying in the set $\tilde{E}_0$). Next we compute coefficients numbered 97,20,19, then 32,13,26, then 102,31, and 12,25. Finally, the coefficients numbered 68,69 are computed by standard smoothness conditions. □

Example 4.2. *The space $\mathcal{S}_4(\triangle_Q)$ is the subspace of $\mathcal{S}_{13}^{5,6}(\triangle_Q) \cap C^8(v_Q)$ that satisfies the twelve special smoothness conditions corresponding to $\{\tau_{8,e_i}^6, \tau_{8,e_i}^7\}_{i=1}^4$ and $\tau_{9,e_1}^6, \tau_{9,e_1}^7, \tau_{9,e_1}^8, \tau_{9,e_1}^9$.*

Discussion. The dimension of $\mathcal{S}_4(\triangle_Q)$ is 152, and the MDS for this macroelement is shown in Figure 3. It consists of 28 points in each of the disks $D_3(v_i)$ (marked with crosses) and 10 points corresponding to item (2) of Theorem 3.1 for each edge of Q (marked with triangles). □

5. The case $r = 2m + 1$.

Theorem 5.1. *Given $r = 2m + 1$, let $\mathcal{S}_r(\triangle_Q)$ be the linear subspace of all splines s in $\mathcal{S}_{6m+3}^{2m+1,3m+1}(\triangle_Q)$ that satisfy the following set of additional smoothness conditions:*

$$s \in C^{4m+1}(v_Q), \tag{5.1}$$

$$\tau_{3m+i+2,e_l}^{2m+1+j} s = 0, \qquad 1 \le j \le 2i, \quad 1 \le i \le m-1, \quad l = 1,2,3,4, \tag{5.2}$$

$$\tau_{4m+2,e_l}^{2m+1+j} s = 0, \qquad 1 \le j \le 2m, \quad l = 1,2,3. \tag{5.3}$$

Then

$$\dim \mathcal{S}_r(\triangle_Q) = 26m^2 + 42m + 16.$$

Moreover, the following set $\mathcal{M}_r$ of domain points is an MDS:

(1) $D^{T^{[i]}}_{3m+1}(v_i)$ *for* $i = 1, 2, 3, 4,$

(2) $\{\xi^{T^{[i]}}_{j,3m+1,3m-j+2}, \dots, \xi^{T^{[i]}}_{j,3m-j+2,3m+1}\}$ *for* $j = 1, \dots, 2m+1$ *and* $i = 1, 2, 3, 4.$

Proof. First, we show that $\mathcal{M}_r$ is a determining set. Suppose that $s \in \mathcal{S}_r(\triangle_Q)$ and that we have set the coefficients of s corresponding to all domain points in $\mathcal{M}_r$. Then, using the usual smoothness conditions, we solve for the unset coefficients corresponding to domain points in the disks $D_{3m+1}(v_i)$ for $i = 1, 2, 3, 4$.

Next we use Lemma 2.1 to compute the coefficients corresponding to points on the rings $R_{3m+i+2}(v_l)$ for $i = 0, \dots, m-1$ and $l = 1, 2, 3, 4$. On the ring $R_{3m+i+2}(v_l)$, this involves solving a system of $2(m+i)+1$ linear equations. Then we compute coefficients on the rings $R_{4m+2}(v_l)$ for $l = 1, 2, 3$.

Using the lemma, we now compute the $4m+1$ unset coefficients corresponding to the sets E_0 and $\tilde{E}_0$ defined in the proof of Theorem 3.1. The remaining coefficients in $T^{[1]} \cup T^{[2]} \cup T^{[3]}$ are computed by alternately working on the sets E_i and $\tilde{E}_i$ for $i = 1, \dots, r-1$. Finally, we compute the remaining coefficients in $T^{[4]}$ from the C^r smoothness conditions.

We have shown that all coefficients of s are determined by those corresponding to the domain points in the set $\mathcal{M}_r$. This shows that $\mathcal{M}_r$ is a determining set.

To see that $\mathcal{M}_r$ is an MDS, consider $\mathcal{S}^{2m+1}_{6m+3}(\triangle_Q) \cap C^{4m+1}(v_Q)$. By Theorem 2.2 in [18] the dimension of this space is $32m^2 + 46m + 16$. Our space $\mathcal{S}_r(\triangle_Q)$ is the subspace which satisfies the $4m^2 + 2m$ special conditions (5.2)–(5.3) and the supersmoothness $C^{3m+1}(v_i)$ for $i = 1, 2, 3, 4$. Enforcing the supersmoothness requires an additional $2m^2 + 2m$ conditions. Thus

$$\begin{aligned}(32m^2 + 46m + 16) - (4m^2 + 2m) - (2m^2 + 2m)&\\ \le \dim \mathcal{S}_r(\triangle_Q) \le \#\mathcal{M}_r = 26m^2 + 42m + 16.&\end{aligned}$$

Since the expression on the left equals the one on the right, we conclude that it is equal to the dimension of $\mathcal{S}_r(\triangle_Q)$, and $\mathcal{M}_r$ is an MDS. □

6. Examples. In this section, we illustrate the construction of section 5.

EXAMPLE 6.1. *The space $\mathcal{S}_3(\triangle_Q)$ is the subspace of $\mathcal{S}^{3,4}_9(\triangle_Q) \cap C^5(v_Q)$ that satisfies the six special smoothness conditions corresponding to $\tau^4_{6,e_i}, \tau^5_{6,e_i}$ for $i = 1, 2, 3$.*

Discussion. The dimension of $\mathcal{S}_3(\triangle_Q)$ is 84, and the MDS for this macroelement is shown in Figure 4. It consists of 15 points in each of the disks $D_4(v_i)$ (marked with crosses) and 6 points corresponding to item (2) of Theorem 5.1 along each edge of Q (marked with triangles). □

EXAMPLE 6.2. *The space $\mathcal{S}_5(\triangle_Q)$ is the subspace of $\mathcal{S}^{5,7}_{15}(\triangle_Q) \cap C^9(v_Q)$ that satisfies the twenty special smoothness conditions corresponding to $\{\tau^6_{9,e_i}, \tau^7_{9,e_i}\}^4_{i=1}$ and $\{\tau^6_{10,e_i}, \tau^7_{10,e_i}, \tau^8_{10,e_i}, \tau^9_{10,e_i}\}^3_{i=1}$.*

Discussion. The space $\mathcal{S}_5(\triangle_Q)$ has dimension 204, and the MDS for this macroelement is shown in Figure 5. It consists of 36 points in each of the disks $D_5(v_i)$ (marked with crosses) and 15 points corresponding to item (2) in Theorem 5.1 along each edge of Q (marked with triangles). □

7. Superspline spaces with stable bases. Let $\Diamond$ be a quadrangulation of a domain Ω with vertices $\{v_i\}^V_{i=1}$. Suppose E is the number of edges. Let $\diamondplus$ be the triangulation obtained by inserting both diagonals in each quadrilateral Q in $\Diamond$. Let

$$\mathcal{S}_r(\diamondplus) := \{s \in C^r(\Omega) \;:\; s|_Q \in \mathcal{S}_r(\triangle_Q) \text{ for all } Q \in \Diamond\}, \tag{7.1}$$

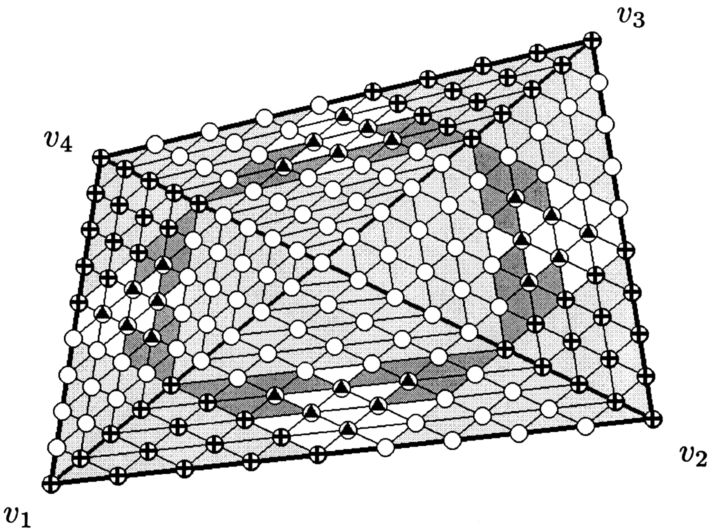


FIG. 4. *The C^3 macroelement.*

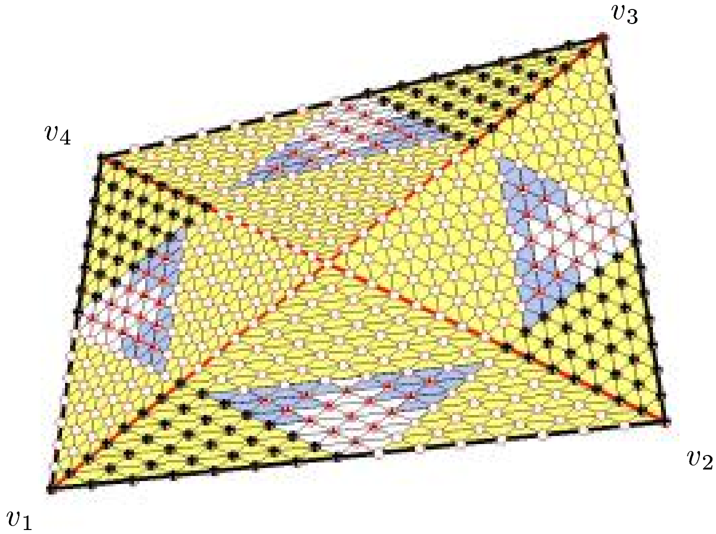


FIG. 5. *The C^5 macroelement.*

where $\mathcal{S}_r(\triangle_Q)$ are the spaces defined in Theorems 3.1 and 5.1. Let

$$d_r = \begin{cases} 6m+1, & r = 2m, \\ 6m+3, & r = 2m+1, \end{cases} \tag{7.2}$$

and

$$\rho_r = \begin{cases} 3m, & r = 2m, \\ 3m+1, & r = 2m+1. \end{cases} \tag{7.3}$$

THEOREM 7.1. *For all $r \geq 1$,*

$$\dim \mathcal{S}_r(\diamondplus) = \binom{\rho_r + 2}{2} V + \binom{r+1}{2} E. \tag{7.4}$$

Moreover, the following set $\mathcal{M}_r$ of domain points forms an MDS:

(1) *For each vertex $v \in \Diamond$, choose $D^T_{\rho_r}(v)$, where T is some triangle in $\diamondplus$ with vertex at v;*

(2) *for each edge $e \in \Diamond$, choose $\{\xi^T_{j,\rho_r,d_r-\rho_r-j}, \ldots, \xi^T_{j,d_r-\rho_r-j,\rho_r}\}$ for $j = 1, \ldots, r$, where T is some triangle in $\diamondplus$ sharing the edge e.*

Proof. First, we show that $\mathcal{M}_r$ is a determining set. For each vertex $v \in \Diamond$, using the smoothness conditions, item (1) determines all coefficients corresponding to points in the disk $D_{\rho_r}(v)$. Similarly, if $\tilde{T}$ is a second triangle sharing the edge e, then item (2) determines the corresponding coefficients in both T and $\tilde{T}$. The claim then follows from Theorems 3.1 and 5.1.

To show that $\mathcal{M}_r$ is an MDS, we now construct the dual basis corresponding to $\mathcal{M}_r$. For each $\xi \in \mathcal{M}_r$, let B_ξ be the unique spline in $\mathcal{S}_r(\diamondplus)$ such that

$$\lambda_\eta B_\xi = \delta_{\xi,\eta}, \qquad \eta \in \mathcal{M}_r, \tag{7.5}$$

where λ_η is the linear functional which picks off the B-coefficient corresponding to the domain point η.

In view of (7.5), the splines in $\mathcal{B} := \{B_\xi\}_{\xi \in \mathcal{M}_r}$ are linearly independent, and thus $\mathcal{B}$ forms a basis for $\mathcal{S}_r(\diamondplus)$. It follows that $\dim \mathcal{S}_r(\diamondplus) = \#\mathcal{M}_r$, which is the number in (7.4). □

It is easy to see that the dual basis functions constructed in the above proof have local support. In particular,

(1) if ξ is a point as in item (1) of Theorem 7.1, then $\operatorname{supp}(B_\xi)$ is contained in the union of all quadrilaterals of $\diamondplus$ sharing the vertex v;

(2) if ξ is a point as in item (2) of Theorem 7.1, then $\operatorname{supp}(B_\xi)$ is contained in $Q \cup \tilde{Q}$, where e is the edge between Q and $\tilde{Q}$. (If e is a boundary edge of a quadrilateral Q, then the support is simply Q.)

LEMMA 7.2. *Let $\{B_\xi\}_{\xi \in \mathcal{M}}$ be the set of dual basis splines constructed in the proof of Theorem 7.1. Then there exists a constant K depending only on the smallest angle in $\diamondplus$ such that $\|B_\xi\| \leq K$ for all $\xi \in \mathcal{M}_r$.*

Proof. Fix $\xi \in \mathcal{M}_r$, and let B_ξ be the corresponding dual basis spline. We examine the size of its B-coefficients. By definition, $c_\xi = 1$ and $c_\eta = 0$ for all other $\eta \in \mathcal{M}_r$. The remaining B-coefficients of B_ξ are computed by using smoothness conditions or solving the linear systems of equations appearing in Lemma 2.1 of [2]. These involve matrices whose inverses are bounded in norm by a constant depending only on the smallest angle in $\diamondplus$. This shows that all of the B-coefficients of B_ξ are bounded by a constant K, and the result follows. □

THEOREM 7.3. *The dual basis $\{B_\xi\}_{\xi \in \mathcal{M}_r}$ is a stable basis in the sense that there exist constants K_1, K_2 depending only on the smallest angle in $\diamondplus$ such that, for all choices of the coefficient vector $c = (c_\xi)_{\xi \in \mathcal{M}_r}$,*

$$K_1 \|c\|_\infty \;\leq\; \Big\| \sum_{\xi \in \mathcal{M}_r} c_\xi B_\xi \Big\|_\infty \;\leq\; K_2 \|c\|_\infty. \tag{7.6}$$

Proof. The proof follows in the same way as the proof of Theorem 2.3 of [7]. □

We conclude this section with an approximation result. Given a function f in $L_1(\Omega)$ and an integer $0 \le k \le d_r$, let

$$Q_k f := \sum_{\xi \in \mathcal{M}_r} \lambda_{\xi,k} f \, B_\xi,$$

where $\lambda_{\xi,k}$ is the linear functional defined in section 10 of [11].

THEOREM 7.4. *Fix $1 \le p \le \infty$. Suppose f lies in the Sobolev space $W_p^{k+1}(\Omega)$ for some $0 \le k \le d_r$. Then*

$$\|D_x^\alpha D_y^\beta (f - Q_k f)\|_p \le K |\diamondplus|^{k+1-\alpha-\beta} |f|_{k+1,p} \tag{7.7}$$

for $0 \le \alpha + \beta \le k$, where $|\diamondplus|$ is the mesh size of $\diamondplus$ (i.e., the diameter of the largest triangle), and $|f|_{k+1,p}$ is the usual Sobolev seminorm. If Ω is convex, then the constant K depends only on d_r, p, k, and the smallest angle in $\diamondplus$. If Ω is nonconvex, it also depends on the Lipschitz constant $L_{\partial\Omega}$ associated with the boundary of Ω.

Proof. The proof follows in the same way as the proof of Theorem 1.1 of [11]. □

8. Remarks.

Remark 8.1. We used the java code described in [1] to check the macroelements described in this paper and to generate the figures. The code can be used or downloaded from http://www.math.utah.edu/~alfeld.

Remark 8.2. Theorem 10.1 of [13] implies that, in order to obtain macroelements on $\triangle_Q$ which will join with C^r smoothness when constructed on the individual quadrilaterals of a quadrangulation, we must require supersmoothness of order ρ_r at the vertices of Q, where ρ_r is defined in (7.3). This implies that we cannot use polynomials of degree lower than the d_r given in (7.2).

Remark 8.3. While there is a unique choice of minimal degree and minimal supersmoothness at the vertices of Q, our choice of extra smoothness conditions is not the only choice which leads to macroelements based on the natural set of degrees of freedom, i.e., other sets of τ's will also work.

Remark 8.4. In view of the connection between derivatives and B-coefficients, Theorem 7.1 immediately implies that, given $f \in C^{\rho_r}(\Omega)$, there exists a unique spline $s \in \mathcal{S}_r(\diamondplus)$ which solves the Hermite interpolation problem

$$D_x^\nu D_y^\mu s(v) = D_x^\nu D_y^\mu f(v), \qquad 0 \le \nu + \mu \le \rho_r, \qquad v \in \Diamond, \tag{8.1}$$

and

$$D_e^j s(\eta_{e,i}^j) = D_e^j f(\eta_{e,i}^j), \qquad 1 \le i \le j, \quad 1 \le j \le r, \tag{8.2}$$

for all edges e of $\Diamond$. Here D_e denotes the perpendicular derivative to the edge $e := \langle u_1, u_2 \rangle$, and

$$\eta_{e,i}^j s := \frac{(j+1-i)u_1 + iu_2}{j+1}, \qquad i = 1, \dots, j.$$

A simple application of the Bramble–Hilbert lemma shows that this interpolant satisfies the error bounds of Theorem 7.4 with $p = \infty$. For error bounds for other bivariate spline spaces, see [4].

Remark 8.5. In view of Remark 8.2, it is clear that the natural set of degrees of freedom for a C^r macroelement on $\triangle_Q$ are precisely the derivative information described in (8.1)–(8.2). For a comparison with the natural degrees of freedom for smooth macroelements defined on Clough–Tocher and Powell–Sabin splits, see [2, 3, 12, 13].

REFERENCES

[1] P. Alfeld, *Bivariate splines and minimal determining sets*, J. Comput. Appl. Math., 119 (2000), pp. 13–27.

[2] P. Alfeld and L. L. Schumaker, *Smooth finite elements based on Clough-Tocher triangle splits*, Numer. Math., to appear.

[3] P. Alfeld and L. L. Schumaker, *Smooth finite elements based on Powell-Sabin triangle splits*, Adv. Comput. Math., to appear.

[4] C. deBoor, *Multivariate piecewise polynomials*, in Acta Numerica, Acta Numer., A. Iserles, ed., Cambridge University Press, Cambridge, UK, 1993, pp. 65–109.

[5] P. G. Ciarlet, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1991, pp. 17–351.

[6] J. F. Ciavaldini and J. C. Nédélec, *Sur l'élément de Fraeijs de Veubeke et Sander*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 8 (1974), pp. 29–46.

[7] O. Davydov and L. L. Schumaker, *On stable local bases for bivariate polynomial spline spaces*, Constr. Approx., 18 (2002), pp. 87–116.

[8] B. Fraeijs de Veubeke, *A conforming finite element for plate bending*, J. Solids Structures, 4 (1968), pp. 95–108.

[9] A. Ibrahim and L. L. Schumaker, *Super spline spaces of smoothness r and degree $d \geq 3r+2$*, Constr. Approx., 7 (1991), pp. 401–423.

[10] M. J. Lai, *On dual functionals of polynomials in B-form*, J. Approx. Theory, 67 (1991), pp. 19–37.

[11] M. J. Lai and L. L. Schumaker, *On the approximation power of bivariate splines*, Adv. Comput. Math., 9 (1998), pp. 251–279.

[12] M. J. Lai and L. L. Schumaker, *Macro-elements and stable local bases for splines on Clough-Tocher triangulations*, Numer. Math., 88 (2001), pp. 105–119.

[13] M. J. Lai and L. L. Schumaker, *Macro-elements and stable local bases for splines on Powell-Sabin triangulations*, Math. Comp., to appear.

[14] M. Laghchim-Lahlou and P. Sablonnière, *Composite quadrilateral finite elements of class C^r*, in Mathematical Methods in Computer Aided Geometric Design, T. Lyche and L. Schumaker, eds., Academic Press, New York, 1989, pp. 313–418.

[15] M. Laghchim-Lahlou and P. Sablonnière, *Quadrilateral finite elements of FVS type and class C^ρ*, Numer. Math., 70 (1995), pp. 229–243.

[16] P. Sablonnière and M. Laghchim-Lahlou, *Elements finis polynomiaux composé de classe C^r*, C. R. Acad. Sci. Paris Sér. I Math., 316 (1993), pp. 503–508.

[17] G. Sander, *Bornes supérieures et inférieures dans l'analyse matricielle des plaques en flexion-torsion*, Bull. Soc. Roy. Sci. Liège, 33 (1964), pp. 456–494.

[18] L. L. Schumaker, *Dual bases for spline spaces on cells*, Comput. Aided Geom. Design, 5 (1988), pp. 277–284.

SIAM J. MATH. ANAL.
Vol. 33, No. 5, pp. 1117–1143

EDGE BIFURCATIONS FOR NEAR INTEGRABLE SYSTEMS VIA EVANS FUNCTION TECHNIQUES*

TODD KAPITULA† AND BJÖRN SANDSTEDE‡

Abstract. When studying the linear stability of waves for near integrable systems, a fundamental problem is the location of the point spectrum of the linearized operator. Internal modes may be created upon the perturbation, i.e., eigenvalues may bifurcate out of the continuous spectrum, even if the corresponding eigenfunction is not initially localized. This phenomenon is also known as an *edge bifurcation*. It has recently been shown that the Evans function is a powerful tool when one wishes to detect an edge bifurcation and track the resulting eigenvalues. It has been an open question as to the role played by the solutions to the Lax pair, associated with the integrable problem, in the construction of the Evans function and the detection of edge bifurcations. Using the Zakharov–Shabat eigenvalue problem and the massive Thirring model as illustrations, we show the connection between the inverse scattering formalism and the linear stability analysis of waves. In particular, we show a direct connection between the scattering coefficients and the Evans function. Last, using perturbations of the massive Thirring model, we show how the Evans function can be used to predict the location of bifurcating edge eigenvalues.

Key words. travelling waves, stability, Evans function, inverse scattering theory

AMS subject classifications. 35P05, 37K10, 35P25

PII. S0036141000372301

1. Introduction.

Much work has been done recently on the detection of edge bifurcations, or internal modes, for perturbed integrable systems (see, for instance, [15, 16, 17, 25, 27, 31]). When considering the stability of solitons for an integrable system, it is typically a straightforward problem to locate the spectrum associated with the operator which arises from linearizing about the soliton. This is possible primarily due to the fact that the integrable problem has so much structure. The basic issue is then to try to determine the spectrum for the perturbed problem when this structure is no longer available.

In order to illustrate the issues involved, consider the following well-studied example, the perturbed focusing nonlinear Schrödinger equation (NLS):

$$\mathrm{i}q_t + \frac{1}{2}q_{xx} - \omega q + |q|^2 q = \mathrm{i}\epsilon R(q, q^*) \qquad (+\text{complex conjugate (c.c.)}),$$

where $q(x,t) \in \mathbb{C}$ for any $(x,t) \in \mathbb{R} \times \mathbb{R}^+$ and where $\omega \in \mathbb{R}^+$ and $0 \leq \epsilon \leq 1$. When $\epsilon = 0$, the NLS has the soliton solution

$$\phi(x,\omega) = \sqrt{2\omega}\,\mathrm{sech}(\sqrt{2\omega}\,x).$$

After linearizing the unperturbed NLS about ϕ, we obtain

$$\mathrm{i}q_t + \frac{1}{2}q_{xx} - \omega q + \phi^2(2q + q^*) = 0 \qquad (+\text{c.c.}),$$

[*] Received by the editors May 15, 2000; accepted for publication (in revised form) September 18, 2001; published electronically February 14, 2002.
http://www.siam.org/journals/sima/33-5/37230.html

[†] Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131 (kapitula@math.unm.edu). This author was partially supported by National Science Foundation grant DMS-9803408.

[‡] Department of Mathematics, Ohio State University, Columbus, OH 43210 (sandstede.1@osu.edu). This author was partially supported by National Science Foundation grant DMS-9971703 and by an Alfred P. Sloan Research Fellowship.

and the associated eigenvalue problem is therefore

$$\mathrm{i}\lambda q + \frac{1}{2}q_{xx} - \omega q + \phi^2(2q + q^*) = 0. \qquad (+\text{c.c.})$$

Upon computing the spectrum of the linearized operator, one finds that $\lambda = 0$ is an isolated eigenvalue of geometric multiplicity two and algebraic multiplicity four, and the rest of the spectrum is a continuous spectrum which resides on the imaginary axis in the bands $|\operatorname{Im}\lambda| \geq \omega$ [21, 39]. Thus, when $\epsilon = 0$, the wave is linearly stable in that there exists no unstable spectrum (i.e., no spectral elements λ with $\operatorname{Re}\lambda > 0$). Here, to fix notation, we divide the spectrum of a given operator into two disjoint sets. The point spectrum consists of all isolated eigenvalues with finite multiplicity, while the essential (or continuous) spectrum is the complement of the point spectrum in the spectrum.

A fundamental question is: If the wave persists, what happens to its linear stability for $\epsilon > 0$? There are two issues to consider. The first is the fate of the point eigenvalues at $\lambda = 0$. Depending on the type of perturbation and what invariances of the NLS are preserved under the perturbation, some of these eigenvalues will generally move. This issue comes up in general in the study of perturbed Hamiltonian systems (see [14] and references therein). The perturbed eigenvalues can be tracked, for instance, by using the Evans function [13, 14] or via a Lyapunov-Schmidt reduction. There has also been a great deal of formal work using adiabatic and variational ideas. (A small subset is [3, 21, 40, 26, 35].) The idea behind these methods is that the effect of radiation on the evolution of the perturbed wave is of higher order and can thus be neglected at first order in a perturbation expansion. For the focusing NLS, the rigorous eigenvalue analysis and formal work are in agreement; however, this is not always the case, for it has recently been shown in [15] that the results disagree for the defocusing NLS.

The second issue, and the primary focus of this paper, is to locate the rest of the spectrum. The location of the continuous spectrum is straightforward (see Henry [11]), and one typically assumes that the perturbation is such that the continuous spectrum remains in the closed left-half plane. A more troublesome problem is determining the location of the rest of the point spectrum. Under the perturbation of the PDE, it is possible for point eigenvalues to move out of the continuous spectrum. If this occurs, then one says that an edge bifurcation has occurred or an internal mode has been created. It must be emphasized that an edge bifurcation can happen even if the corresponding eigenfunctions for the unperturbed problem are not localized. Indeed, the recent work of [16, 17, 25, 31] has shown this to be the case for the perturbed NLS. For the NLS, it turns out that an edge bifurcation can happen only at the edge of the continuous spectrum, i.e., at the points $\lambda = \pm\mathrm{i}\omega$; furthermore, at most one eigenvalue can pop out of the continuous spectrum at each of these two points. A priori, however, this fact is not obvious, as it appears to be possible for an edge bifurcation to occur along any point in the continuous spectrum: since it is known that the eigenfunctions associated with the continuous spectrum are all bounded but nondecaying, it seems to be plausible that a small perturbation of the PDE can cause these bounded eigenfunctions to decay as $|x| \to \infty$ and therefore to become true eigenfunctions.

This idea naturally leads to the question: What is the underlying mechanism that determines where an edge bifurcation may take place? We investigate this question for integrable PDEs. Thus consider a Lax pair

$$\mathbf{v}_x = L\mathbf{v}, \qquad \mathbf{v}_t = M\mathbf{v},$$

where $\mathbf{v} \in \mathbb{C}^n$, and L and M are 2×2 matrices that depend on (x, t). The associated integrable PDE

$$\mathbf{u}_t = \mathcal{K}(\mathbf{u}) \tag{1.1}$$

is derived from the compatibility condition $\mathbf{v}_{xt} = \mathbf{v}_{tx}$, i.e., from $L_t - M_x + [L, M] = 0$. The direct scattering problem consists of finding the solutions to $\mathbf{v}_x = L\mathbf{v}$ that are oscillatory as $|x| \to \infty$. These solutions are called Jost functions, and they can be parametrized by their asymptotic spatial wavenumber, or frequency, k for $x \to -\infty$. Associated with the Jost function with wavenumber k is the transmission coefficient $a(k)$, which measures the amplitude of the Fourier mode with wavenumber k as $x \to \infty$. Roughly speaking,

$$\phi(k, x) \sim \mathrm{e}^{-\mathrm{i}kx} \text{ as } x \to -\infty, \qquad \phi(k, x) \sim a(k)\mathrm{e}^{-\mathrm{i}kx} \text{ as } x \to \infty,$$

where $\phi(k, x)$ is the Jost function with wavenumber k.

We are interested in finding solutions to the eigenvalue problem which arises from linearizing (1.1) about the soliton. It turns out that certain quadratic combinations of the Jost functions, the so-called adjoint squared eigenfunctions, satisfy this linearized problem [21, 22]. Thus one can recover asymptotically oscillatory solutions to the linearized problem which correspond to eigenmodes in the continuous spectrum. Furthermore, if one knows the transmission coefficient $a(k)$, then one can determine the precise behavior of these solutions as $|x| \to \infty$. Of course, not all of the solutions here will be known; however, as will be seen, this will be unimportant as long as one has sufficient information about $a(k)$. As we shall see, the transmission coefficient actually encodes all the information needed to predict edge bifurcations.

We use the Evans function, $E(\lambda)$, to study edge bifurcations. The Evans function $E(\lambda)$ is an analytic function of the spectral parameter λ which has the property that $E(\lambda) = 0$ if and only if λ is a point eigenvalue; furthermore, the order of the zero is the algebraic multiplicity of the eigenvalue [4, 6, 7, 8, 9, 30]. Originally, the Evans function was defined only away from the continuous spectrum. We showed in [16, 17] (see also [10]) that the Evans function can actually be extended across the essential spectrum. Edge bifurcations manifest themselves as zeros of the Evans function that move out of the continuous spectrum. Thus, by locating the zeros of $E(\lambda)$ embedded in the continuous spectrum, one can determine where an edge bifurcation will take place and can also track these zeros under the perturbation.

In order to finish the problem, one must then be able to calculate the Evans function on the continuous spectrum. For the class of problems under consideration in this paper, this calculation is possible. It is found that for integrable PDEs arising from the Zakharov–Shabat eigenvalue problem, the Evans function restricted to the continuous spectrum vanishes precisely at branch points of the continuous spectrum and at points where the transmission coefficient $a(k)$ vanishes for some real k (see section 3.3 for details). For simplicity, we exclude the second possibility (see Assumption 2.1) and focus instead on edge bifurcations at branch points. In this situation, the total number of bifurcating eigenvalues depends on the order of the branch point (see section 3 for details). To illustrate the effects of perturbations, we consider the massive Thirring model, where an edge bifurcation can happen at four branch points—two are at the edge of the continuous spectrum, and two are contained within the continuous spectrum. The reason for this phenomenon is that the massive Thirring model possesses *two* dispersion relationships which are used to describe the continuous spectrum, while the PDEs arising from the Zakharov–Shabat eigenvalue problem contain only one. We then utilize the resulting Evans function to present a perturbation expansion for the bifurcating eigenvalues.

Pelinovsky and Sulem [32, 33] have recently considered a problem which is complementary to that presented herein. They consider an integrable PDE (such as the focusing NLS) which possesses a soliton solution and ask the question about the evolution of small perturbations of this soliton. Since the system is integrable, the question can be answered by performing a detailed analysis of the associated scattering problem. Using a different technique than that presented herein, they study edge bifurcations of simple eigenvalues for the scattering problem (and not for the integrable PDE itself as we do). They show that if an edge bifurcation occurs, then a certain matching condition must hold. (Note, however, that their approach does not appear to prove the converse.) In addition, they show that an edge bifurcation leads to the creation of new solitons.

The paper is organized as follows. In section 2, we discuss the Zakharov–Shabat eigenvalue problem as the prototypical example. In section 3, we show how the squared eigenfunctions of the scattering problem are related to the linear stability problem of the underlying integrable PDE. Furthermore, we calculate the Evans function for the integrable PDE. Last, in section 4, we use the massive Thirring model to illustrate what happens when several branch points collide and to present a perturbation analysis of the Evans function that predicts the location of eigenvalues that move out of the continuous spectrum upon adding perturbation to the PDE.

NOTATION 1.1. *Throughout this paper, we denote the complex conjugate of a complex number q by q^*; we emphasize that $\bar{q}$ does not denote the complex conjugate of q.*

2. Direct and inverse scattering for the Zakharov–Shabat problem.

Much of the following discussion follows that given in [2, 24] (see also [1]) and is included to set up the notation and for the sake of completeness.

2.1. The evolution equation.

We begin by choosing complex-valued functions r and q that depend on (x,t) such that $\mathbf{u}(x,t) = (r,q)(x,t)$ decays to zero exponentially as $|x| \to \infty$ for every t. Consider the 2×2 scattering problem given by

$$\mathbf{v}_x = \begin{pmatrix} -\mathrm{i}k & q(x,t) \\ r(x,t) & \mathrm{i}k \end{pmatrix} \mathbf{v}, \tag{2.1}$$

where $\mathbf{v} = (v_1, v_2)$ again depends on (x,t). We assume that the time-dependence of $\mathbf{v}$ is governed by the linear equation

$$\mathbf{v}_t = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \mathbf{v}, \tag{2.2}$$

where A, B, C, and D are complex-valued functions that depend on $\mathbf{u} = (r,q)$ and on k. The compatibility condition $\mathbf{v}_{tx} = \mathbf{v}_{xt}$ is satisfied provided we have that $D = -A$ and

$$\begin{aligned} A_x + rB - qC &= 0, \\ B_x - q_t + 2Aq + 2\mathrm{i}kB &= 0, \\ C_x - r_t - 2Ar - 2\mathrm{i}kC &= 0. \end{aligned} \tag{2.3}$$

One way of solving the compatibility condition (2.3) is to assume an ansatz of the form

$$A = \Omega(k) + \sum_{j=0}^{n-1} A_j k^j, \qquad B = \sum_{j=0}^{n-1} B_j k^j, \qquad C = \sum_{j=0}^{n-1} C_j k^j,$$

so that A, B, and C are polynomial functions of the eigenvalue parameter k for a given dispersion relation

$$\Omega(k) = \sum_{j=0}^{n} d_j k^j. \tag{2.4}$$

Here the coefficients d_j are given complex numbers, while the coefficients A_j, B_j, and C_j depend on (r, q) in such a fashion that they decay to zero as $|x| \to \infty$. Substituting the above ansatz into (2.3), solving recursively starting at $j = n - 1$, and working down to $j = 0$, we obtain

$$A_j = \int_{-\infty}^{x} (qC_j - rB_j)\, \mathrm{d}y, \qquad \begin{pmatrix} C_j \\ B_j \end{pmatrix} = \sum_{\ell=j+1}^{n} \mathrm{i} d_\ell (\mathcal{L}^A(\mathbf{u}))^{\ell-1-j} \mathbf{u} \tag{2.5}$$

for $j = 0, \ldots, n - 1$. The operator $\mathcal{L}^A(\mathbf{u})$ that appears in (2.5) arises naturally when solving (2.3) inductively and is given by

$$\mathcal{L}^A(\mathbf{u})\mathbf{v} := -\frac{1}{2}\mathrm{i}\sigma_3 \partial_x \mathbf{v} + \mathrm{i}\mathbf{u} \int_{-\infty}^{x} (qv_1 - rv_2)\, \mathrm{d}y, \tag{2.6}$$

where $\mathbf{u} = (r, q)$, $\mathbf{v} = (v_1, v_2)$, and σ_3 is the Pauli spin matrix

$$\sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

In addition, as part of the compatibility condition on the level $j = 0$, we obtain the evolution equation

$$\sigma_3 \mathbf{u}_t + 2\Omega(\mathcal{L}^A(\mathbf{u}))\mathbf{u} = 0 \tag{2.7}$$

for $\mathbf{u}(x, t) = (r, q)(x, t)$.

In summary, for any given polynomial dispersion relation $\Omega(k)$, we can construct the evolution equation (2.7) by means of the Zakharov–Shabat scattering problem (2.1) and (2.2). The class of equations that can be realized by (2.7), upon using an appropriate dispersion relation $\Omega(k)$, includes KdV, mKdV, and NLS (see [2, p. 258]). Direct and inverse scattering theory allow us to actually solve (2.7).

2.2. The direct scattering problem. First, we describe how the scattering data are obtained from the functions (r, q). This map involves certain solutions, the so-called Jost functions, to (2.1). For any k with $\operatorname{Im} k > 0$, there are unique solutions $\phi(k, x)$ and $\psi(k, x)$ to (2.1) that satisfy

$$\lim_{x \to -\infty} \phi(k, x)\mathrm{e}^{\mathrm{i}kx} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \lim_{x \to \infty} \psi(k, x)\mathrm{e}^{-\mathrm{i}kx} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

The solutions $\mathrm{e}^{\mathrm{i}kx}\phi(k, x)$ and $\mathrm{e}^{-\mathrm{i}kx}\psi(k, x)$ are analytic in k for $\operatorname{Im} k > 0$. We emphasize that these solutions also depend on t; we will, however, suppress this dependence in our notation. The first piece of the scattering data, the transmission coefficient $a(k)$, is given by

$$a(k) := \phi(k,x) \wedge \psi(k,x) := \det \begin{pmatrix} \phi_1(k,x) & \psi_1(k,x) \\ \phi_2(k,x) & \psi_2(k,x) \end{pmatrix}.$$

Note that we have

$$\lim_{x\to\infty} \phi(k,x)\mathrm{e}^{\mathrm{i}kx} = a(k)\begin{pmatrix}1\\0\end{pmatrix}$$

for $\operatorname{Im} k > 0$ since the trace of the matrix in (2.1) is zero. It is desirable to analytically extend the transmission coefficient $a(k)$ across the real axis $\operatorname{Im} k = 0$. This extension has been carried out in [2, p. 268] under the assumption that the potentials $(r,q)(x,t)$ decay exponentially to zero as $|x| \to \infty$. The gap lemma [10, 17] deals with this issue in more generality. Roughly speaking, in this context, it states that if the system (2.1) approaches a constant system exponentially fast as $|x| \to \infty$, which is guaranteed by the assumption on r and q, then the solutions ϕ and ψ can be analytically extended. After extending the coefficient $a(k)$, we have that, when $k \in \mathbb{R}$,

$$\lim_{x\to\infty} \left[\phi(k,x) - a(k)\mathrm{e}^{-\mathrm{i}kx}\begin{pmatrix}1\\0\end{pmatrix} - b(k)\mathrm{e}^{\mathrm{i}kx}\begin{pmatrix}0\\1\end{pmatrix}\right] = 0$$

for some function $b(k)$, the so-called reflection coefficient, which may not be analytic in k. In a similar fashion, one can construct solutions $\bar\phi(k,x)$ and $\bar\psi(k,x)$ to (2.1) for $\operatorname{Im} k < 0$ that satisfy

$$\lim_{x\to-\infty} \bar\phi(k,x)\mathrm{e}^{-\mathrm{i}kx} = \begin{pmatrix}0\\-1\end{pmatrix}, \qquad \lim_{x\to\infty} \bar\psi(k,x)\mathrm{e}^{\mathrm{i}kx} = \begin{pmatrix}1\\0\end{pmatrix}.$$

Recall that $\bar\phi$ does not denote the complex conjugate of ϕ. The above solutions are analytic in k for $\operatorname{Im} k < 0$ and can again be extended across $\operatorname{Im} k = 0$ in an analytic fashion. The associated transmission coefficient $\bar a(k)$ is given by

$$\bar a(k) = \bar\phi(k,x) \wedge \bar\psi(k,x)$$

and is analytic for $\operatorname{Im} k \le 0$. For $\operatorname{Im} k < 0$, we have

$$\lim_{x\to\infty} \bar\phi(k,x)\mathrm{e}^{-\mathrm{i}kx} = \bar a(k)\begin{pmatrix}0\\-1\end{pmatrix},$$

while, for $\operatorname{Im} k = 0$, we have

$$\lim_{x\to\infty} \left[\bar\phi(k,x) - \bar a(k)\mathrm{e}^{\mathrm{i}kx}\begin{pmatrix}0\\-1\end{pmatrix} - \bar b(k)\mathrm{e}^{-\mathrm{i}kx}\begin{pmatrix}1\\0\end{pmatrix}\right] = 0$$

for some function $\bar b(k)$ that may not be analytic.

2.3. The adjoint squared eigenfunctions. Next we use the Jost functions to define the squared eigenfunctions $\Psi(k,x)$ and $\bar\Psi(k,x)$ as well as the adjoint squared eigenfunctions $\Psi^A(k,x)$ and $\bar\Psi^A(k,x)$. The functions Ψ and $\bar\Psi$ are given by

$$\Psi(k,x) = \begin{pmatrix}\psi_1(k,x)^2\\ \psi_2(k,x)^2\end{pmatrix}, \qquad \bar\Psi(k,x) = \begin{pmatrix}\bar\psi_1(k,x)^2\\ \bar\psi_2(k,x)^2\end{pmatrix}$$

for $\operatorname{Im} k \geq 0$ and $\operatorname{Im} k \leq 0$, respectively, while Ψ^A and $\bar{\Psi}^A$ are defined via

$$\Psi^A(k,x) = \begin{pmatrix} \phi_2(k,x)^2 \\ -\phi_1(k,x)^2 \end{pmatrix}, \qquad \bar{\Psi}^A(k,x) = \begin{pmatrix} \bar{\phi}_2(k,x)^2 \\ -\bar{\phi}_1(k,x)^2 \end{pmatrix}$$

again for $\operatorname{Im} k \geq 0$ and $\operatorname{Im} k \leq 0$, respectively. All of these functions are analytic in k in their domain of definition. The adjoint squared eigenfunctions satisfy

$$\begin{aligned} &\lim_{x\to-\infty} \Psi^A(k,x)\mathrm{e}^{2\mathrm{i}kx} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \qquad \lim_{x\to\infty} \Psi^A(k,x)\mathrm{e}^{2\mathrm{i}kx} = a(k)^2 \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \\ &\lim_{x\to-\infty} \bar{\Psi}^A(k,x)\mathrm{e}^{-2\mathrm{i}kx} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \lim_{x\to\infty} \bar{\Psi}^A(k,x)\mathrm{e}^{-2\mathrm{i}kx} = \bar{a}(k)^2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \end{aligned} \tag{2.8}$$

for $\operatorname{Im} k > 0$ and $\operatorname{Im} k < 0$, respectively. If $b(k) = \bar{b}(k) = 0$ for all $k \in \mathbb{R}$, then (2.8) is also true for $\operatorname{Im} k = 0$; otherwise, additional terms that involve the reflection coefficients have to be added in a straightforward fashion using the asymptotics of the Jost functions. We remark that the (adjoint) squared eigenfunctions also depend on t. A straightforward computation that uses (2.1) shows that

$$(\mathcal{L}^A(\mathbf{u}) - k)\Psi^A(k,x) = (\mathcal{L}^A(\mathbf{u}) - k)\bar{\Psi}^A(k,x) = 0$$

for any $k \in \mathbb{R}$. For 1-soliton solutions with $r = -q^*$, the adjoint squared eigenfunctions can be calculated explicitly (see [20, section 3]).

2.4. The inverse scattering problem. In the previous sections, we started with a solution $(r,q)(x,t)$ to (2.7) and then associated with that solution the scattering data, the Jost functions, and eventually the squared eigenfunctions. Inverse scattering allows us to, at least in principle, find solutions to (2.7) for a given set of scattering data. In other words, the inverse scattering problem consists of mapping given scattering data to solutions $\mathbf{u} = (r,q)$ of (2.7). We briefly describe the procedure and refer to [2, section IV.B and Appendix 5] and [20, p. 124] for more details and proofs.

Recall that $\Omega(k)$ denotes the polynomial dispersion relation that we chose earlier. We begin with the so-called primordial scattering data

$$\left\{ \frac{\bar{b}_0(k)}{a_0(k)},\ \frac{b_0(k)}{\bar{a}_0(k)};\ k \in \mathbb{R} \right\}, \qquad \left\{ (k_j, \beta_j)_{j=1,N},\ (\bar{k}_j, \bar{\beta}_j)_{j=1,\bar{N}} \right\} \tag{2.9}$$

consisting of the functions a_0, $\bar{a}_0$, b_0, and $\bar{b}_0$, where we assume that the fractions appearing in (2.9) are bounded uniformly in $k \in \mathbb{R}$ and pairs of complex numbers with $\Omega(k_j) = 0 = \Omega(\bar{k}_j)$ for all j. Thus the primordial scattering data involve fractions of the scattering data. We remark that the scattering data can be reconstructed from the primordial scattering data [2, Appendix 5]. We evolve the primordial scattering data by defining

$$\frac{\bar{b}(k,t)}{a(k,t)} = \mathrm{e}^{2\Omega(k)t}\frac{\bar{b}_0(k)}{a_0(k)}, \qquad \frac{b(k,t)}{\bar{a}(k,t)} = \mathrm{e}^{-2\Omega(k)t}\frac{b_0(k)}{\bar{a}_0(k)}. \tag{2.10}$$

Using these data, one can, at least in principle, reconstruct the Jost functions $\phi(k,x)$ and $\bar{\phi}(k,x)$ by solving certain integral equations [2, section IV.B]. Using the Jost functions, we obtain the adjoint squared eigenfunctions and from these a solution

$(r,q)(x,t)$ of (2.7) via

$$\begin{aligned}\begin{pmatrix} r \\ q \end{pmatrix} = \frac{1}{\pi}\int_{-\infty}^{\infty} \left[\frac{\bar{b}(k)}{a(k)}\Psi^A(k,x) + \frac{b(k)}{\bar{a}(k)}\bar{\Psi}^A(k,x)\right] \mathrm{d}k \\ -2\mathrm{i}\sum_{j=1}^{N}\beta_j\Psi^A(k_j,x) + 2\mathrm{i}\sum_{j=1}^{\bar{N}}\bar{\beta}_j\bar{\Psi}^A(\bar{k}_j,x).\end{aligned}$$

In the next section, we restrict our analysis to transmission coefficients $a(k)$ and $\bar{a}(k)$ that are nonzero for $k \in \mathbb{R}$ and that have a finite number of zeros off the real axis $\operatorname{Im} k = 0$. For transmission coefficients with these properties, it was shown by Kaup [19] that the adjoint squared eigenfunctions formed by the Jost functions ϕ and $\bar{\phi}$ form a basis for $L^2(\mathbb{R},\mathbb{C}) \cap L^1(\mathbb{R},\mathbb{C})$.

2.5. Spectral stability of 1-solitons. We consider 1-solitons which are solutions to (2.7) whose scattering data are rather simple [1, 2]. The precise requirements on the scattering data of 1-solitons are summarized in the following assumption.

Assumption 2.1. Assume that $\mathbf{u}_0$ is a stationary, i.e., time-independent, 1-soliton. More precisely, we assume that the following is true: we have $N = \bar{N} = 1$ in (2.9), the associated reflection coefficients $b(k)$ and $\bar{b}(k)$ are identically zero for $k \in \mathbb{R}$, and the associated transmission coefficients $a(k)$ and $\bar{a}(k)$ are nonzero for any $k \in \mathbb{R}$ and have each precisely one simple zero off the real axis $\operatorname{Im} k = 0$.

We remark that the simple zeros of $a(k)$ and $\bar{a}(k)$ are given by $k = k_1$ and $k = \bar{k}_1$, respectively, which are defined in (2.9) [2, Appendix 5].

Note that 1-solitons typically arise as rotating waves $\mathbf{u}(x,t) = \mathrm{e}^{\mathrm{i}\omega t}\mathbf{u}_0(x)$ or as travelling waves $\mathbf{u}(x,t) = \mathbf{u}_0(x - ct)$ (or a combination of both) for suitable real numbers ω and c. Upon changing the coefficients d_0 and d_1 of the dispersion relation $\Omega(k)$ in (2.4), we can always arrange to have $\omega = 0$ and $c = 0$. The change in the dispersion relation amounts to transforming (2.7) into a corotating and comoving coordinate frame (see [2, (2.7)] for an explicit example).

We are interested in the stability of soliton solutions to (2.7) upon adding perturbations that destroy the integrability of (2.7). To investigate this question, we need to have information about the spectrum and the associated eigenfunctions of the integrable PDE (2.7) linearized about a soliton. Thus define

$$\mathcal{K}(\mathbf{u}) := -2\sigma_3\Omega(\mathcal{L}^A(\mathbf{u}))\mathbf{u} \tag{2.11}$$

so that (2.7) is given by $\mathbf{u}_t = \mathcal{K}(\mathbf{u})$ and such that its linearization about the soliton $\mathbf{u}_0$ is

$$\mathbf{v}_t = \mathcal{K}'(\mathbf{u}_0)\mathbf{v}. \tag{2.12}$$

Upon seeking solutions to (2.12) of the form $\mathbf{v}(x,t) = \mathrm{e}^{\lambda t}\mathbf{v}(x)$, we get the associated eigenvalue problem

$$[\mathcal{K}'(\mathbf{u}_0) - \lambda]\mathbf{v} = 0. \tag{2.13}$$

We then have the following lemma.

Lemma 2.2. *Suppose that Assumption* 2.1 *is met. The adjoint squared eigenfunctions* $\Psi^A(k,x)$ *and* $\bar{\Psi}^A(k,x)$ *then satisfy*

$$\begin{aligned} &[\mathcal{K}'(\mathbf{u}_0) - 2\Omega(k)]\sigma_3\Psi^A(k,x) = 0 \quad \textit{for any } k \textit{ with } \operatorname{Im} k \geq 0, \\ &[\mathcal{K}'(\mathbf{u}_0) + 2\Omega(k)]\sigma_3\bar{\Psi}^A(k,x) = 0 \quad \textit{for any } k \textit{ with } \operatorname{Im} k \leq 0. \end{aligned} \tag{2.14}$$

In other words, $\sigma_3\Psi^A(k,x)$ and $\sigma_3\bar{\Psi}^A(k,x)$ are eigenfunctions of the linear operator $\mathcal{K}'(\mathbf{u}_0)$. We remark that the adjoint squared eigenfunctions do not depend on t, since we assumed that the underlying solution $\mathbf{u}_0 = (r_0, q_0)$ is time-independent.

Proof. The claim is a consequence of results by Kaup and Newell (see [24, Appendix B] or [29, section 6.4]). The idea is to compute the variation of (2.7) with respect to one of the primordial scattering data. Hence we fix $\ell \in \mathbb{R}$ and replace the original scattering data by the fractions

$$(1+\epsilon\delta(k-\ell))\frac{\bar{b}(k)}{a(k)},$$

where δ denotes the δ-distribution and ϵ is a small perturbation parameter. The following arguments are formal but can be made rigorous by approximating the δ-distribution with localized bump functions. Upon calculating the derivative of (2.7) with respect to ϵ, we conclude that

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}\begin{pmatrix} r_t \\ q_t \end{pmatrix} = \mathcal{K}'(\mathbf{u}_0)\frac{\mathrm{d}}{\mathrm{d}\epsilon}\begin{pmatrix} r \\ q \end{pmatrix}$$

so that

$$\sigma_3\frac{\mathrm{d}}{\mathrm{d}\epsilon}\begin{pmatrix} r_t \\ -q_t \end{pmatrix} = \mathcal{K}'(\mathbf{u}_0)\sigma_3\frac{\mathrm{d}}{\mathrm{d}\epsilon}\begin{pmatrix} r \\ -q \end{pmatrix}. \tag{2.15}$$

From [29, (6.55)], we know that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\epsilon}\begin{pmatrix} r \\ -q \end{pmatrix} &= -\frac{1}{\pi}\int_{-\infty}^{\infty}\left[\frac{\mathrm{d}}{\mathrm{d}\epsilon}\left(\frac{\bar{b}(k)}{a(k)}\right)\Psi^A(k,x) - \frac{\mathrm{d}}{\mathrm{d}\epsilon}\left(\frac{b(k)}{\bar{a}(k)}\right)\bar{\Psi}^A(k,x)\right]\mathrm{d}k \\ &\quad +2\mathrm{i}\frac{\mathrm{d}\beta_1}{\mathrm{d}\epsilon}\Psi^A(k_1,x) + 2\mathrm{i}\frac{\mathrm{d}\bar{\beta}_1}{\mathrm{d}\epsilon}\bar{\Psi}^A(\bar{k}_1,x).\end{aligned}$$

Upon taking the time derivative of this expression and using (2.10), we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\epsilon}\begin{pmatrix} r_t \\ -q_t \end{pmatrix} &= -\frac{1}{\pi}\int_{-\infty}^{\infty}\left[\frac{\mathrm{d}}{\mathrm{d}\epsilon}\left(\frac{\bar{b}(k)}{a(k)}\right)(2\Omega(k)\Psi^A(k,x) + \partial_t\Psi^A(k,x))\right. \\ &\quad \left. -\frac{\mathrm{d}}{\mathrm{d}\epsilon}\left(\frac{b(k)}{\bar{a}(k)}\right)(-2\Omega(k)\bar{\Psi}^A(k,x) + \partial_t\bar{\Psi}^A(k,x))\right]\mathrm{d}k \\ &\quad +2\mathrm{i}\frac{\mathrm{d}\beta_1}{\mathrm{d}\epsilon}\partial_t\Psi^A(k_1,x) + 2\mathrm{i}\frac{\mathrm{d}\bar{\beta}_1}{\mathrm{d}\epsilon}\partial_t\bar{\Psi}^A(\bar{k}_1,x).\end{aligned}$$

Evaluating these expressions at $b = \bar{b} = 0$ and using that the adjoint squared eigenfunctions of the 1-soliton do not depend on time, we get

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}\begin{pmatrix} r \\ -q \end{pmatrix} = -\frac{1}{\pi}\Psi^A(\ell,x), \qquad \frac{\mathrm{d}}{\mathrm{d}\epsilon}\begin{pmatrix} r_t \\ -q_t \end{pmatrix} = -\frac{2\Omega(\ell)}{\pi}\Psi^A(\ell,x).$$

Substituting this result into (2.15), we obtain

$$2\Omega(\ell)\sigma_3\Psi^A(\ell,x) = \mathcal{K}'(\mathbf{u}_0)\sigma_3\Psi^A(\ell,x),$$

which is the desired result. Using $(1+\epsilon\delta(k-\ell))b(k)/\bar{a}(k)$ and following the steps above, we see that

$$-2\Omega(\ell)\sigma_3\bar{\Psi}^A(\ell,x) = \mathcal{K}'(\mathbf{u}_0)\sigma_3\bar{\Psi}^A(\ell,x)$$

for any $\ell \in \mathbb{R}$. Since the adjoint squared eigenfunctions Ψ^A and the dispersion relation are analytic in ℓ, we conclude that Ψ^A satisfies the above equation for any ℓ with $\operatorname{Im} \ell \geq 0$. This completes the argument. □

The following lemma shows that the only element in the point spectrum Σ_{pt} of the linearized operator $\mathcal{K}'(\mathbf{u}_0)$ is $\lambda = 0$.

LEMMA 2.3. *Suppose that Assumption* 2.1 *is met. If* $\mathbf{v}(x)$ *is a solution of* (2.13) *such that* $\mathbf{v}(x)$ *decays to zero exponentially as* $|x| \to \infty$*, then* $\lambda = 0$ *or* $\mathbf{v} = 0$.

Proof. Suppose that $\mathbf{v}(x)$ is a solution of (2.13) such that $\mathbf{v}(x)$ decays to zero exponentially as $|x| \to \infty$. Owing to the completeness result in [19], we then have

$$\begin{aligned}(2.16)\quad \sigma_3 \mathbf{v}(x) = \int_{-\infty}^{\infty} &\left[\alpha(k)\Psi^A(k,x) + \bar{\alpha}(k)\bar{\Psi}^A(k,x)\right] \mathrm{d}k \\ &+\beta\Psi^A(k_1,x) + \bar{\beta}\bar{\Psi}^A(\bar{k}_1,x) + \gamma\frac{\mathrm{d}}{\mathrm{d}k}\Psi^A(k_1,x) + \bar{\gamma}\frac{\mathrm{d}}{\mathrm{d}k}\bar{\Psi}^A(\bar{k}_1,x)\end{aligned}$$

for certain exponentially decaying functions $\alpha(k), \bar{\alpha}(k)$ and complex numbers $\beta, \bar{\beta}, \gamma$, and $\bar{\gamma}$. Here, the derivatives of Ψ^A and $\bar{\Psi}^A$ with respect to k, evaluated at k_1, are generalized eigenfunctions of $\mathcal{K}'(\mathbf{u}_0)$, belonging to the eigenvalue $\lambda = 0$, that need to be added to the set introduced in section 2.3 to make it complete [19]. Multiplying (2.16) by σ_3, applying the operator $[\mathcal{K}'(\mathbf{u}_0) - \lambda]$ to both sides of the above equation, and exploiting the properties of the adjoint squared eigenfunctions established in [19] and in Lemma 2.2, we obtain

$$\begin{aligned}0 = \int_{-\infty}^{\infty} &\left[(2\Omega(k) - \lambda)\alpha(k)\Psi^A(k,x) - (2\Omega(k) + \lambda)\bar{\alpha}(k)\bar{\Psi}^A(k,x)\right] \mathrm{d}k \\ &+\lambda(\gamma - \beta)\Psi^A(k_1,x) + \lambda(\bar{\gamma} - \bar{\beta})\bar{\Psi}^A(\bar{k}_1,x) - \lambda\gamma\frac{\mathrm{d}\Psi^A}{\mathrm{d}k}(k_1,x) - \lambda\bar{\gamma}\frac{\mathrm{d}\bar{\Psi}^A}{\mathrm{d}k}(\bar{k}_1,x).\end{aligned}$$

We conclude that necessarily $\alpha(k) = \bar{\alpha}(k) = 0$ for all k. Furthermore, we have either $\lambda = 0$ or the coefficients $\beta, \bar{\beta}, \gamma$, and $\bar{\gamma}$ vanish. In the latter case, we have $\mathbf{v} = 0$. □

Note that, as a consequence of the results in [19], $\lambda = 0$ has geometric multiplicity two, and each of the geometric eigenvalues admits a maximal Jordan block of length two. In fact, the eigenfunctions are given by $\Psi^A(k_1,x)$ and $\bar{\Psi}^A(\bar{k}_1,x)$, where k_1 and $\bar{k}_1$ are defined in (2.9), and the associated generalized eigenfunctions are $\frac{\mathrm{d}}{\mathrm{d}k}\Psi^A(k_1,x)$ and $\frac{\mathrm{d}}{\mathrm{d}k}\bar{\Psi}^A(\bar{k}_1,x)$, respectively. This can be confirmed by taking derivatives of (2.14) with respect to k, evaluated at $k = k_1$ and $k = \bar{k}_1$, using that $\Omega(k_1) = \Omega(\bar{k}_1) = 0$, and exploiting (2.8) together with the assumption that $k = k_1$ and $k = \bar{k}_1$ are simple zeros of $a(k)$ and $\bar{a}(k)$, respectively.

It is straightforward to compute the essential spectrum Σ_{ess} of $\mathcal{K}'(\mathbf{u}_0)$. The operator $\mathcal{K}'(\mathbf{u}_0)$ is a relatively compact perturbation of the linearization $\mathcal{K}'(0)$ about the asymptotic rest state $\mathbf{u} = 0$ which is given by

$$\mathcal{K}'(0) = -2\sigma_3\Omega\left(-\frac{1}{2}\mathrm{i}\sigma_3\partial_x\right)$$

on account of (2.6) and (2.11). As a consequence, the essential spectrum of $\mathcal{K}'(\mathbf{u}_0)$ is

$$(2.17)\qquad \Sigma_{\mathrm{ess}} = \{\lambda \in \mathbb{C};\ \lambda = \pm 2\Omega(k) \text{ for some } k \in \mathbb{R}\}.$$

More precisely, the radiation modes are given by $\exp(\lambda t + \mathrm{i}kx)\mathbf{v}_0$, where λ and k are related via $\lambda = 2\Omega(-k/2)$ for $\mathbf{v}_0 = (1,0)$ and $\lambda = -2\Omega(k/2)$ for $\mathbf{v}_0 = (0,1)$. In particular, the essential spectrum is (marginally) stable, i.e., lies on the imaginary axis, if and only if the coefficients in (2.7) are purely imaginary. Since the soliton is certainly unstable if the essential spectrum is unstable, we assume henceforth that the above condition on the coefficients is met.

Assumption 2.4. We assume that the coefficients d_j of the dispersion relation $\Omega(k)$, introduced in (2.4), are purely imaginary; i.e., we have $d_j \in \mathrm{i}\mathbb{R}$ for $j = 0, \ldots, n$.

3. The Evans function.

Now that those aspects of the inverse scattering formalism that will be necessary for the following analysis have been addressed, we will proceed to show how Assumption 2.1 yields precise information as to how one can locate those points in the continuous spectrum at which discrete eigenvalues may move out upon adding a perturbation to the integrable PDE (2.7). The tool we choose to use is the Evans function, $E(\lambda)$. Thus, consider the eigenvalue problem (2.13)

$$[\mathcal{K}'(\mathbf{u}_0) - \lambda]\mathbf{v} = 0. \tag{3.1}$$

We assume that this eigenvalue problem can be rewritten as the first-order system

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbf{Y} = [M(\lambda) + R(x)]\mathbf{Y}, \qquad \mathbf{Y}(x) \in \mathbb{C}^{2n}, \tag{3.2}$$

where n is the degree of the polynomial dispersion relation (2.4) and therefore of the linearized PDE (3.1). The matrix $R(x)$ converges to zero exponentially fast as $|x| \to \infty$. Note that $\mathbf{v}$ satisfies (3.1) if and only if $Y = (1, \frac{\mathrm{d}}{\mathrm{d}x}, \ldots, \frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}})\mathbf{v}$ satisfies (3.2).

3.1. The construction of the Evans function.

We briefly recall the construction of the Evans function following the approach in [36] that avoids differential forms and refer to [4] for the original approach via differential forms. We begin by defining the Evans function for values of $\lambda \in \mathbb{C}$ such that λ is not in the essential spectrum given in (2.17). For such values of λ, the matrix $M(\lambda)$ is hyperbolic, i.e., its spectrum contains no points on the imaginary axis. We are interested in locating isolated eigenvalues which correspond to those values of λ for which (3.2) or, equivalently, (3.1) has a nonzero localized solution. We accomplish this by constructing all solutions to (3.2) that decay as $x \to -\infty$ and all solutions that decay as $x \to \infty$. The associated initial data at a given value of x then define two subspaces such that λ is an eigenvalue if and only if those two subspaces have a nontrivial intersection, leading to a solution of (3.2) that decays as $x \to \pm\infty$. Therefore, assuming that $M(\lambda)$ has n_u eigenvalues with positive real part and $n_s = 2n - n_u$ eigenvalues with negative real part, choose two sets $\{\mathbf{Y}_j(\lambda, x)\}_{j=1,\ldots,n_u}$ and $\{\mathbf{Y}_j(\lambda, x)\}_{j=n_u+1,\ldots,2n}$ of linearly independent solutions of (3.2) with the following properties: the above solutions depend analytically on λ, $\mathbf{Y}_j(\lambda, x)$ converges to zero exponentially as $x \to -\infty$ for $j = 1, \ldots, n_u$, and $\mathbf{Y}_j(\lambda, x)$ converges to zero exponentially as $x \to \infty$ for $j = n_u + 1, \ldots, 2n$. Note that such a choice is always possible (see, e.g., [36]). The Evans function $E(\lambda)$ is the complex-valued function defined by

$$E(\lambda) = \exp\left(-\int_0^x \mathrm{tr}[M(\lambda) + R(s)]\,\mathrm{d}s\right) \det(\mathbf{Y}_1(\lambda, x) \ldots \mathbf{Y}_{2n}(\lambda, x)),$$

where we consider the $2n$ vectors $\mathbf{Y}_j(\lambda, x)$ as column vectors in a $2n \times 2n$ matrix. By construction, λ is in the point spectrum Σ_{pt} if and only if $E(\lambda) = 0$, as this then leads to a solution (the eigenfunction) of (3.2) that decays exponentially as $|x| \to \infty$.

In fact, the algebraic multiplicity of λ, considered as an eigenvalue of $\mathcal{K}'(\mathbf{u}_0)$, is equal to the order of λ, considered as a zero of $E(\lambda)$; see [4]. As we shall see now, the set $\{\mathbf{Y}_j(\lambda,x)\}_{j=1,\ldots,n_\mathrm{u}}$ can be constructed from the adjoint squared eigenfunctions $\sigma_3\Psi^A(k,x)$ and $\sigma_3\bar{\Psi}^A(k,x)$.

To see this, fix again λ with $\operatorname{Re}\lambda \geq 0$ such that $\lambda \notin \Sigma_{\mathrm{ess}}$. For any such fixed λ, solve the equation $\lambda = 2\Omega(k)$ subject to the constraint $\operatorname{Im} k > 0$, and denote the corresponding roots by $k_1, \ldots, k_\ell$, counted with their order.[1] Analogously, solve $\lambda = -2\Omega(k)$ subject to the constraint $\operatorname{Im} k < 0$, and denote the solutions by $\bar{k}_1, \ldots, \bar{k}_{\bar{\ell}}$, again counted with their order. Note that ℓ and $\bar{\ell}$ are independent of λ, since ℓ and $\bar{\ell}$ can change only when one of the k's becomes real, which can happen only for $\lambda \in \Sigma_{\mathrm{ess}}$. We exclude those values of λ, however. We remark that the solutions $k_1, \ldots, k_\ell$ and $\bar{k}_1, \ldots, \bar{k}_{\bar{\ell}}$ depend continuously on λ.

The roots k_j and $\bar{k}_j$ of the dispersion relation and the eigenvalues ν_j and $\bar{\nu}_j$ of the matrix $M(\lambda)$ are related via

$$\nu_j = -2\mathrm{i}k_j, \qquad \bar{\nu}_j = 2\mathrm{i}\bar{k}_j \tag{3.3}$$

for $j = 1, \ldots, n$. Note that the unstable eigenvalues of $M(\lambda)$ are those that correspond to $k_1, \ldots, k_\ell$ and $\bar{k}_1, \ldots, \bar{k}_{\bar{\ell}}$, so that $\ell + \bar{\ell} = n_u$.

Consider the eigenvalue problem (3.1). Suppose first that λ is such that $k_1, \ldots, k_\ell$ are ℓ distinct numbers, while $\bar{k}_1, \ldots, \bar{k}_{\bar{\ell}}$ are $\bar{\ell}$ distinct numbers. As a consequence of Lemma 2.2, $\sigma_3\Psi^A(k_j,x)$ for $j = 1, \ldots, \ell$ and $\sigma_3\bar{\Psi}^A(\bar{k}_j,x)$ for $j = 1, \ldots, \bar{\ell}$ are solutions to the eigenvalue problem (3.1). Furthermore, on account of (2.8) and the above discussion, these solutions are linearly independent, they decay exponentially as $x \to -\infty$, and any solution to (3.1) with this property is, in fact, a linear combination of the above adjoint squared eigenfunctions. Again, due to (2.8), any linear combination of the $\Psi^A(k_1,x), \ldots, \Psi^A(k_\ell,x)$ and $\bar{\Psi}^A(\bar{k}_1,x), \ldots, \bar{\Psi}^A(\bar{k}_{\bar{\ell}},x)$ grows exponentially as $x \to \infty$ except for those λ for which $a(k_j)$ or $\bar{a}(\bar{k}_j)$ vanishes for some j. Assumption 2.1 precludes this except when $\lambda = 0$.

Next suppose that one of the $k_1, \ldots, k_\ell$ has order $m > 1$ as a zero of $\lambda = 2\Omega(k)$. In other words, assume that, possibly after relabeling the roots, $k_1 = \cdots = k_m$ and $k_j \neq k_1$ for $j = m+1, \ldots, \ell$, so that

$$\frac{\mathrm{d}^j}{\mathrm{d}k^j}\Omega(k)\Big|_{k=k_1, j=1,\ldots,m-1} = 0, \qquad \frac{\mathrm{d}^m}{\mathrm{d}k^m}\Omega(k)\Big|_{k=k_1} \neq 0.$$

Differentiating (2.14) and (2.8) with respect to k shows that $\sigma_3 \frac{\mathrm{d}^j}{\mathrm{d}k^j}\Psi^A(k_1,x)$ satisfies (3.1) with

$$\lim_{x\to-\infty} \frac{\mathrm{e}^{2\mathrm{i}k_1 x}}{(-2\mathrm{i}x)^j} \frac{\mathrm{d}^j\Psi^A}{\mathrm{d}k^j}(k_1,x) = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad \lim_{x\to\infty} \frac{\mathrm{e}^{2\mathrm{i}k_1 x}}{(-2\mathrm{i}x)^j} \frac{\mathrm{d}^j\Psi^A}{\mathrm{d}k^j}(k_1,x) = a(k_1)^2 \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

for $j = 0, \ldots, m-1$. Furthermore, these solutions together with the remaining functions $\Psi^A(k_{m+1},x), \ldots, \Psi^A(k_\ell,x)$ and $\bar{\Psi}^A(\bar{k}_1,x), \ldots, \bar{\Psi}^A(\bar{k}_{\bar{\ell}},x)$ form a linearly independent set of solutions upon multiplying them by σ_3. Again, linear combinations of these functions generate all solutions to (3.1) that decay as $x \to -\infty$, and any such combination grows exponentially as $x \to \infty$ except when $\lambda = 0$.

The above discussion provides a different proof of Lemma 2.2. The crucial assumption is that $a(k)\bar{a}(k) \neq 0$ except for those k for which $\Omega(k) = 0$ so that $\lambda = 0$.

[1] Note that we abuse notation: the roots $k_1, \ldots, k_\ell$ are not related to the numbers appearing in (2.9).

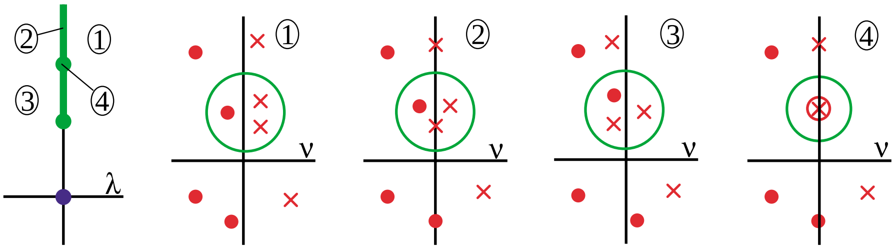


FIG. 3.1. *The leftmost plot shows the spectrum of the integrable PDE in the complex λ-plane with the upper bullet labeled 4 being the branch point λ_*. The other four plots, numbered 1–4, show the eigenvalues ν and $\bar{\nu}$ (related via (3.3) to the roots k and $\bar{k}$ of the dispersion relation) of the matrix $M(\lambda)$ for λ to the right (1), on (2), and to the left (3) of the essential spectrum. The inset with label 4 shows these eigenvalues at the branch point, where the three spatial eigenvalues ν inside the circle collapse. In the plots labeled 1–4, bullets and crosses denote the eigenvalues in the stable and unstable sets.*

An extension of the above calculation will allow us to address the issue of locating edge-bifurcation points.

In the last step, we concentrate on the essential spectrum. The idea is to analytically extend the solutions $\mathbf{Y}_j(\lambda, x)$ that we constructed above for $\lambda \notin \Sigma_{\text{ess}}$ into the essential spectrum. Hence again fix λ with $\operatorname{Re}\lambda \geq 0$ such that $\lambda \notin \Sigma_{\text{ess}}$. Define the sets

$$K^u(\lambda) := \{k_1(\lambda), \dots, k_\ell(\lambda)\}, \qquad \bar{K}^u(\lambda) := \{\bar{k}_1(\lambda), \dots, \bar{k}_{\bar{\ell}}(\lambda)\},$$

which depend continuously on λ. Analogously, denote by $k_{\ell+1}, \dots, k_n$ the roots of $\lambda = 2\Omega(k)$ that satisfy $\operatorname{Im} k < 0$ and by $\bar{k}_{\bar{\ell}+1}, \dots, \bar{k}_n$ the roots of $\lambda = -2\Omega(k)$ with $\operatorname{Im} k > 0$, and define

$$K^s(\lambda) := \{k_{\ell+1}(\lambda), \dots, k_n(\lambda)\}, \qquad \bar{K}^s(\lambda) := \{\bar{k}_{\bar{\ell}+1}(\lambda), \dots, \bar{k}_n(\lambda)\}.$$

We then continue these sets of roots into the essential spectrum, which is possible since the roots k of $\lambda = 2\Omega(k)$ and $\lambda = -2\Omega(k)$ depend continuously on λ. Note that the sets $K^u(\lambda)$ and $K^s(\lambda)$ have a nonempty intersection precisely when $\lambda = 2\Omega(k)$ and $\Omega'(k) = 0$ for some $k \in \mathbb{R}$. In fact, $K^u(\lambda) \cap K^s(\lambda)$ consists exactly of all elements k with the above property. The analogous statement is, of course, true for $\bar{K}^u(\lambda)$ and $\bar{K}^s(\lambda)$. We say that λ is a branch point of the essential spectrum if $K^u(\lambda) \cap K^s(\lambda)$ or $\bar{K}^u(\lambda) \cap \bar{K}^s(\lambda)$ are nonempty. This is also illustrated in Figure 3.1.

On account of [17] and the above discussion, the Evans function $E(\lambda)$ vanishes for $\lambda \in \Sigma_{\text{ess}}$ if and only if there are coefficients $b_j, \bar{b}_j \in \mathbb{C}$, with $j = 1, \dots, n$ such that the function

$$v(x) = \sum_{j=1}^{\ell} b_j \Psi^A(k_j(\lambda), x) + \sum_{j=1}^{\bar{\ell}} \bar{b}_j \bar{\Psi}^A(\bar{k}_j(\lambda), x) \tag{3.4}$$

has the asymptotic behavior

$$v(x) = \sum_{j=\ell+1}^{n} b_j \mathrm{e}^{-2\mathrm{i}k_j(\lambda)x} \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \sum_{j=\bar{\ell}+1}^{n} \bar{b}_j \mathrm{e}^{-2\mathrm{i}\bar{k}_j(\lambda)x} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \mathrm{O}(x^{-1}) \tag{3.5}$$

as $x \to \infty$. Of course, at least one of the coefficients b_j and $\bar{b}_j$ has to be nonzero. In other words, we seek solutions to (3.1) or, equivalently, (3.2) that have wavenumbers in K^u or $\bar{K}^u$ as $x \to -\infty$ and wavenumbers in K^s or $\bar{K}^s$ as $x \to \infty$. Comparing (3.4) and (3.5) with (2.8), we see that this occurs exactly when $a(k)\bar{a}(k) = 0$ for some $k \in \mathbb{C}$ with $\lambda = 2\Omega(k)$ or $\lambda = -2\Omega(k)$ or else if λ is a branch point of the essential spectrum. Indeed, if $k \in K^u(\lambda) \cap K^s(\lambda)$, then $k = k_1(\lambda) = k_{\ell+1}(\lambda)$, possibly after relabeling the roots, and $v(x) = \Psi^A(k, x)$ satisfies (3.5).

In summary, we have proved the following theorem.

Theorem 3.1. *Assume that Assumptions* 2.1 *and* 2.4 *are met. The Evans function $E(\lambda)$ vanishes only at $\lambda = 0$ and at points λ for which there is a $k \in \mathbb{R}$ such that $\lambda = 2\Omega(k)$ (or $\lambda = -2\Omega(k)$) and $\Omega'(k) = 0$.*

Recall our original motivation: we wanted to locate all points in the essential spectrum where discrete eigenvalues may move out upon adding a perturbation to the original PDE. Any such point corresponds to a zero of the Evans function, and the above theorem indeed gives all zeros of the Evans function: apart from $\lambda = 0$, these zeros are in one-to-one correspondence with the branch points of the essential spectrum, i.e., with any point λ such that there is a $k \in \mathbb{R}$ with $\lambda = 2\Omega(k)$, or $\lambda = -2\Omega(k)$ and $\Omega'(k) = 0$. The remaining issues are to predict how many eigenvalues may move out of the essential spectrum and to locate those eigenvalues for a given perturbation. In the next section, we discuss the first issue.

3.2. The Evans function at branch points.

We want to compute the Evans function $E(\lambda)$ for λ close to a fixed branch point $\lambda_* \in \Sigma_{\text{ess}}$. We assume that λ_* is a branch point of $\lambda = 2\Omega(k)$; the case of branch points of $\lambda = -2\Omega(k)$ can, of course, be handled in the same fashion. Thus we assume that there is a number $k_* \in \mathbb{R}$ such that $\lambda_* = 2\Omega(k_*)$ and $\Omega'(k_*) = 0$. In addition, we assume that the following nondegeneracy condition is met. At the end of this section, we shall comment on the situation where the following assumption is not met.

Assumption 3.2. Assume that λ_* is a nondegenerate branch point, i.e., we have $\Omega'(k) \neq 0$ for any $k \in \mathbb{R}$ such that $\lambda_* = -2\Omega(k)$, or else $\lambda_* = 2\Omega(k)$ with $k \neq k_*$.

Throughout this section, we assume that λ is close to λ_*. Denote by $m \in \mathbb{N}$ the order of k_* as a root of $\lambda_* = 2\Omega(k)$. Note that $m > 1$. As a consequence of Assumption 3.2, we therefore have

$$\lambda = 2\Omega(k) = \lambda_* + (k - k_*)^m \tilde{\Omega}(k), \qquad \tilde{\Omega}(k_*) \in \mathrm{i}\mathbb{R} \setminus \{0\}. \tag{3.6}$$

In particular, if we continue the m roots of (3.6) near k_*, then, upon tracing out one revolution on a circle in λ that is centered at λ_*, these roots undergo a cyclic permutation of length m. After m full revolutions, the roots are labeled as before [18, section II.1.2]. Following [15, 16, 17], we will wish to define the Evans function on an appropriate Riemann surface. We therefore set

$$\lambda = \lambda_* + \gamma^m, \tag{3.7}$$

so that $\arg(\gamma) \in [-\pi/2m, \pi/2m)$ corresponds to the principal sheet of the Riemann surface. The m roots of (3.6) near k_* are then given by

$$k_* + \gamma \exp\left(\frac{2\pi \mathrm{i} j}{m}\right)(\tilde{\Omega}(k_*)^{\frac{1}{m}} + \mathrm{O}(\gamma)), \qquad j = 0, \ldots, m-1. \tag{3.8}$$

For $\operatorname{Re}\lambda > 0$, i.e., for $\arg(\gamma) \in (-\pi/2m, \pi/2m)$, none of the roots in (3.8) is real. (Otherwise, λ would be in the continuous spectrum.) Thus there are numbers m_u and $m_s = m - m_u$ such that m_u of the roots k in (3.8) have $\operatorname{Im} k > 0$ and m_s of them have $\operatorname{Im} k < 0$. We denote these roots by $k_1(\gamma), \dots, k_{m_u}(\gamma)$ and $k_{\ell+1}(\gamma), \dots, k_{\ell+m_s}(\gamma)$, where ℓ is as in section 3.1. Note that these roots are analytic in γ.

The remaining $n-m$ roots k of (3.6) that are not close to k_* can also be divided into roots with $\operatorname{Im} k > 0$ and $\operatorname{Im} k < 0$. We denote those roots by $k_{m_u+1}(\gamma), \dots, k_\ell(\gamma)$ and $k_{\ell+m_s+1}(\gamma), \dots, k_n(\gamma)$, respectively. The roots $\bar{k}_1, \dots, \bar{k}_{\bar{\ell}}$ and $\bar{k}_{\bar{\ell}+1}, \dots, \bar{k}_n$ of $\lambda = -2\Omega(k)$ with $\operatorname{Im} k < 0$ and $\operatorname{Im} k > 0$, respectively, are defined as in section 3.1. As before, the collections of stable and unstable roots define the sets $K^u(\gamma)$ and $K^s(\gamma)$ as well as $\bar{K}^u(\gamma)$ and $\bar{K}^s(\gamma)$.

The key is that these sets are well defined for any γ close to zero. Indeed, we had seen earlier that the sets K^u and K^s (or $\bar{K}^u$ and $\bar{K}^s$) have a nonempty intersection only at branch points with the elements in the intersection given by the roots of the dispersion relation that cause the branch points. Due to Assumption 3.2, those roots are the ones of (3.6) near k_*. These critical roots, however, depend analytically on the new spectral parameter γ.

Figure 3.1 illustrates the situation. Plotted there is the spectrum of the operator $\mathcal{K}'(\mathbf{u}_0)$ and the spectra of the matrix $M(\lambda)$ that appears in (3.2). Note that the eigenvalues of $M(\lambda)$ are related via (3.3) to the roots of the dispersion relation.

The Evans function $E(\gamma)$ is now defined as before except that we count the eigenspaces associated with the roots $k_1(\gamma), \dots, k_\ell(\gamma)$ and $\bar{k}_1(\gamma), \dots, \bar{k}_{\bar{\ell}}(\gamma)$ as unstable and the eigenvalues associated with $k_{\ell+1}, \dots, k_n$ and $\bar{k}_{\bar{\ell}+1}, \dots, \bar{k}_n$ as stable. Again, linearly independent solutions to (3.1) and (3.2) can be computed as in the previous section using the adjoint squared eigenfunctions. The crucial contribution comes from the critical roots $k_1(\gamma), \dots, k_{m_u}(\gamma)$ and $k_{\ell+1}(\gamma), \dots, k_{\ell+m_s}(\gamma)$. First, note that the adjoint squared eigenfunctions $\Psi^A(k_1(\gamma), x), \dots, \Psi^A(k_{m_u}(\gamma), x)$ and $\Psi^A(k_{\ell+1}(\gamma), x), \dots, \Psi^A(k_{\ell+m_s}(\gamma), x)$ are well defined and analytic in γ due to the analytic extension of the Jost function ϕ across $\operatorname{Im} k = 0$. We can therefore use the adjoint squared eigenfunctions to construct solutions of (3.2) in the unstable and stable eigenspaces via

$$\mathbf{Y}_j(x) = \left(1, \frac{\mathrm{d}}{\mathrm{d}x}, \dots, \frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}\right) \Psi^A(k_j(\gamma), x).$$

Note, however, that the functions $\mathbf{Y}_1, \dots, \mathbf{Y}_{m_u}$, and likewise $\mathbf{Y}_{\ell+1}, \dots, \mathbf{Y}_{\ell+m_s}$, are linearly independent only for $\gamma \neq 0$. At $\gamma = 0$, all of these functions coincide. It is certainly possible to construct from the $\mathbf{Y}_j$ a set of linearly independent solutions. Before we do that, note that, due to (3.8), we have

$$\det \begin{pmatrix} 1 & \dots & 1 & 1 & \dots & 1 \\ k_1 & \dots & k_{m_u} & k_{\ell+1} & \dots & k_{\ell+m_s} \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ k_1^{m-1} & \dots & k_{m_u}^{m-1} & k_{\ell+1}^{m-1} & \dots & k_{\ell+m_s}^{m-1} \end{pmatrix} = C_{\mathrm{vmd}} \gamma^{m(m-1)/2}(1 + \mathrm{O}(\gamma)),$$

where

$$C_{\mathrm{vmd}} = \tilde{\Omega}(k_*)^{\frac{m-1}{2}} \prod_{j>l} \left[\exp\left(\frac{2\pi \mathrm{i} j}{m}\right) - \exp\left(\frac{2\pi \mathrm{i} l}{m}\right)\right] \neq 0$$

is a nonzero Vandermonde determinant. Thus, upon exploiting once more the asymptotics (2.8) of the squared eigenfunctions together with the expression (3.8), we see that the minor of the determinant of the matrix with columns $\mathbf{Y}_1(x), \dots, \mathbf{Y}_{m_u}(x)$ is given by $C_u\gamma^{m_u(m_u-1)/2}$ as $x \to \infty$ for some nonzero number C_u. Analogously, the minor of the determinant of the matrix with columns $\mathbf{Y}_{\ell+1}(x), \dots, \mathbf{Y}_{\ell+m_s}(x)$ is given by $C_s\gamma^{m_s(m_s-1)/2}$ as $x \to \infty$ for some $C_s \neq 0$. Thus we may still use the solutions $\mathbf{Y}_1, \dots, \mathbf{Y}_{m_u}$ and $\mathbf{Y}_{\ell+1}, \dots, \mathbf{Y}_{\ell+m_s}$ but must divide the determinant in the definition of the Evans function by the product $C_uC_s\gamma^{m_u(m_u-1)/2}\gamma^{m_s(m_s-1)/2}$ to account for the degeneracy of these solutions. Thus we can finally calculate the Evans function. We obtain

$$\begin{aligned} E(\gamma) &= \exp\left(-\int_0^x \operatorname{tr}[M(\lambda)+R(s)]\,\mathrm{d}s\right)\det(\mathbf{Y}_1(\gamma,x)\dots\mathbf{Y}_{2n}(\gamma,x)) \\ &= \tilde{C}_*(1+\mathrm{O}(\gamma))\gamma^{-m_u(m_u-1)/2}\gamma^{-m_s(m_s-1)/2} \\ &\quad \times \det\begin{pmatrix} 1 & \dots & 1 & 1 & \dots & 1 \\ k_1 & \dots & k_{m_u} & k_{\ell+1} & \dots & k_{\ell+m_s} \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ k_1^{m-1} & \dots & k_{m_u}^{m-1} & k_{\ell+1}^{m-1} & \dots & k_{\ell+m_s}^{m-1} \end{pmatrix} \\ &= C_*(1+\mathrm{O}(\gamma))\gamma^{-m_u(m_u-1)/2}\gamma^{-m_s(m_s-1)/2}\gamma^{m(m-1)/2} \\ &= C_*(1+\mathrm{O}(\gamma))\gamma^{m_um_s}, \end{aligned} \tag{3.9}$$

where the coefficient $C_* \in \mathbb{C}$ is a nonzero number that involves a product of transmission coefficients. Note that we used in (3.9) that the only nonzero contribution to the determinant is from the critical solutions that we constructed above. Indeed, the analysis in section 3.1 applies to all the other solutions. Thus we proved the following theorem.

Theorem 3.3. *Suppose that Assumptions* 2.1, 2.4, *and* 3.2 *are met. Recall the definitions of* m_u *and* $m_s = m - m_u$ *introduced after* (3.8)*. The Evans function* $E(\gamma)$ *is then given by*

$$E(\gamma) = C_*\gamma^{m_um_s} + \mathrm{O}(\gamma^{m_um_s+1}),$$

where $C_* \neq 0$ *and* $\lambda = \lambda_* + \gamma^m$*. Only roots of the Evans function* $E(\gamma)$ *that satisfy* $-\pi/2m < \arg(\gamma) < \pi/2m$ *correspond to eigenvalues of the operator* $\mathcal{K}'(\mathbf{u}_0)$.

As a consequence of Theorem 3.3, upon adding a perturbation of the integrable PDE, there will be m_um_s zeros of the Evans function near $\lambda = \lambda_*$ on the Riemann surface. It is then of interest to understand how many of these zeros correspond to true eigenvalues of the linearized PDE operator and if any oscillatory instabilities arise as a result of the perturbation. This computation involves a Taylor expansion of the Evans function with respect to γ and the perturbation parameter ϵ. The generic case $m = 2$ (with $m_u = m_s = 1$) was considered by Kivshar et al. [25] for conservative perturbations and independently by Kapitula and Sandstede [16, 17] for arbitrary perturbations. Additional applications can be found [27]. Combining the results in [13] and [36], one can derive Taylor expansions of the Evans function also for $m > 1$. Since the expansion depends strongly on m and m_u, we decided not to give any details pertaining to the Zakharov–Shabat problem, and we refer instead to section 4 for an example that involves a branch point of higher order for the massive Thirring model.

3.3. Discussion. We briefly comment on degenerate branch points that violate Assumption 3.2. In this case, there would be several wavenumbers $k_*^{(1)}, \dots, k_*^{(l)} \in \mathbb{R}$ with associated orders $m^{(1)}, \dots, m^{(l)}$ for some $l > 1$. It is then tempting to guess that $E(\gamma) \sim \prod_{j=1}^{l} \gamma^{m_u^{(j)}m_s^{(j)}}$. This is *not* the case, however. The reason is that we need to find the correct scaling for γ in (3.7) that allows us to extend the Evans function to an appropriate Riemann surface. The $m^{(j)}$ roots associated with $k_*^{(j)}$ form a permutation cycle of length $m^{(j)}$. Thus, in (3.7), we need to choose m to be the least common multiple of the numbers $m^{(1)}, \dots, m^{(l)}$. Proceeding as in this section, we see that the order of the Evans functions at degenerate branch points is typically larger than our guess predicts. On the other hand, since only zeros with $-\pi/2m < \arg(\gamma) < \pi/2m$ correspond to true eigenvalues, we also see that less of these zeros are actually meaningful for the original eigenvalue problem.

Last, note that the proof of Theorem 3.1 uses, in a crucial way, the hypothesis that the transmission coefficients $a(k)$ and $\bar{a}(k)$ do not vanish for $k \in \mathbb{R}$ (see Assumption 2.1). In fact, if this assumption is violated so that $a(k_*) = 0$ for some $k_* \in \mathbb{R}$, then the Evans function $E(\lambda)$ vanishes also for $\lambda = 2\Omega(k_*) \in \Sigma_{\text{ess}}$, which may not be a branch point.

4. The massive Thirring model. In the previous two sections, a class of problems was considered in which the PDE had a single dispersion relation associated with it. In this section, a PDE which has two dispersion relations will be discussed. Each will generate a single branch point of the Evans function; furthermore, near each branch point, $E(\gamma) \sim \gamma$ on the appropriate Riemann surface. As a consequence, when doing perturbation expansions, the work presented in [15, 17] will directly apply. However, as a certain parameter is varied, the two branch points will collide so that now $E(\gamma) \sim \gamma^2$. We will perform a perturbation expansion in this case and show that the collision of the branch points can induce an oscillatory instability for the underlying wave.

A light pulse that propagates in a nonlinear grating with a quadratic nonlinearity may generate pulses at higher frequencies via second-harmonic generation. In the case of a longitudinal periodic variation of the linear susceptibility which couples two linearly polarized envelopes at the fundamental frequency, a second-harmonic field is generated through type II second-harmonic generation. The normalized equations which describe this phenomenon are given by

$$\begin{aligned} &\mathrm{i}\partial_t a_1 + \mathrm{i}\partial_x a_1 - \beta a_1 + a_2 + a_3 a_2^* = 0, \\ &\mathrm{i}\partial_t a_2 - \mathrm{i}\partial_x a_2 + a_1 + \beta a_2 + a_3 a_1^* = 0, \\ &\mathrm{i}\partial_t a_3 + \mathrm{i}v\partial_x a_3 + \delta k\, a_3 + a_1 a_2 = 0 \end{aligned} \tag{4.1}$$

(see Trillo [38]). In the above equation, a_j represents the normalized polarized envelope, δk is the second-harmonic generation wave-vector mismatch, and β represents the coupling of the grating to the second-harmonic beam. In the derivation of (4.1), $|\beta| \ll 1$ has been assumed. The parameter v represents the walk-off of the generated second harmonic: when $v = 0$, the second-harmonic velocity is identical with the group velocity, at the fundamental frequency, of the material. Here, we consider only the case $v = 0$ that occurs when the second-harmonic velocity is close to the fundamental frequency of the material [38].

Note that (4.1) is equivariant with respect to phase rotations

$$(a_1, a_2, a_3) \longmapsto (a_1 \mathrm{e}^{\mathrm{i}\alpha}, a_2 \mathrm{e}^{\mathrm{i}\alpha}, a_3 \mathrm{e}^{2\mathrm{i}\alpha}), \qquad \alpha \in \mathbb{R}.$$

Thus we seek rotating waves of the form

$$a_1 = \tilde{a}_1 \mathrm{e}^{-\mathrm{i}\beta x}, \qquad a_2 = \tilde{a}_2 \mathrm{e}^{-\mathrm{i}\beta x}, \qquad a_3 = \tilde{a}_3 \mathrm{e}^{-2\mathrm{i}\beta x}.$$

Substituting this ansatz into (4.1) and dropping the tildes, we obtain

$$\begin{aligned} \mathrm{i}(\partial_t+\partial_x)a_1+a_2+a_3a_2^*&=0,\\ \mathrm{i}(\partial_t-\partial_x)a_2+a_1+a_3a_1^*&=0,\\ \mathrm{i}\partial_t a_3+\delta k\, a_3+a_1a_2&=0. \end{aligned} \tag{4.2}$$

(Recall that we set $v=0$.) Note that (4.2) is Hamiltonian

$$\mathrm{i}\partial_\xi a_j=\frac{\delta H}{\delta a_j^*}\qquad (j=1,2,3),$$

where the Hamiltonian is given by [37]

$$\begin{aligned} H=\int_{-\infty}^{\infty}\Bigg[&\frac{\mathrm{i}}{2}(a_2\partial_x a_2^*-a_1\partial_x a_1^*)+\text{c.c.}\\ &-\delta k|a_3|^2-a_1a_2^*-a_1^*a_2-a_1^*a_2^*a_3-a_1a_2a_3^*\Bigg]\,\mathrm{d}x. \end{aligned}$$

We assume that the mismatch is large and negative, i.e., that $|\delta k|\gg 1$ and $\delta k<0$. Let $\epsilon=-1/\delta k$ so that $0<\epsilon\ll 1$. Upon defining $u_j=a_j/\sqrt{\epsilon}$ for $j=1,2$, (4.2) becomes

$$\begin{aligned} \mathrm{i}(\partial_t+\partial_x)u_1+u_2+a_3u_2^*&=0,\\ \mathrm{i}(\partial_t-\partial_x)u_2+u_1+a_3u_1^*&=0,\\ \mathrm{i}\epsilon\partial_t a_3-a_3+u_1u_2&=0. \end{aligned}$$

We are interested in stationary solutions of the above PDE. For $0<Q<\pi$, we set

$$\begin{aligned} u_j&=\tilde{u}_j\sqrt{1-2\epsilon\cos Q}\exp(-2\mathrm{i}\cos(Q)t),\qquad j=1,2,\\ a_3&=\tilde{a}_3\sqrt{1-2\epsilon\cos Q}\exp(-4\mathrm{i}\cos(Q)t). \end{aligned}$$

Upon substituting and dropping the tildes, we get

$$\begin{aligned} \mathrm{i}\partial_t u_1+\mathrm{i}\partial_x u_1+u_2+\cos(Q)u_1+a_3u_2^*&=0,\\ \mathrm{i}\partial_t u_2-\mathrm{i}\partial_x u_2+u_1+\cos(Q)u_2+a_3u_1^*&=0,\\ \mathrm{i}\epsilon\partial_t a_3-(1-2\epsilon\cos Q)(a_3-u_1u_2)&=0. \end{aligned} \tag{4.3}$$

4.1. Existence of pulses. We seek localized steady-states of (4.3), i.e., solutions to the ODE

$$\begin{aligned} u_1'&=\ \ \mathrm{i}[u_2+\cos(Q)u_1+a_3u_2^*],\\ u_2'&=-\mathrm{i}[u_1+\cos(Q)u_2+a_3u_1^*],\\ 0&=-\mathrm{i}(1-2\epsilon\cos Q)(a_3-u_1u_2), \end{aligned}$$

where $'=\mathrm{d}/\mathrm{d}x$. In particular, a_3 is slaved to the other two variables via $a_3=u_1u_2$, and we arrive at the reduced equation

$$\begin{aligned} u_1'&=\ \ \mathrm{i}[u_2+\cos(Q)u_1+u_1|u_2|^2],\\ u_2'&=-\mathrm{i}[u_1+\cos(Q)u_2+u_2|u_1|^2]. \end{aligned} \tag{4.4}$$

It is important to note that (4.4) is the steady-state problem associated with the massive Thirring model, which after suitable scalings is given by

$$\begin{aligned} \mathrm{i}\partial_t u_1+\mathrm{i}\partial_x u_1+u_2+(\cos Q+|u_2|^2)u_1&=0,\\ \mathrm{i}\partial_t u_2-\mathrm{i}\partial_x u_2+u_1+(\cos Q+|u_1|^2)u_2&=0. \end{aligned} \tag{4.5}$$

The massive Thirring model is an integrable PDE that describes the propagation of optical gap solitons in a periodically modulated nonlinear fiber [5, 22, 23, 28, 34, 38], and its steady-states are known.

LEMMA 4.1. *With* $\Phi(x) = \sin(Q)\,\mathrm{sech}(x\sin Q - \mathrm{i}Q/2)$*, the function* $(u_1, u_2, a_3) = (\Phi, -\Phi^*, -|\Phi|^2)$ *is a stationary solitary pulse to* (4.3) *for each* $|v| < 1$*.*

4.2. The reduced eigenvalue problem. Now that the existence question has been settled, it is desirable to determine the stability of the pulse found in Lemma 4.1. Recall that the pulse is that of the massive Thirring model. In fact, comparing (4.3) and (4.5), we might expect that the linear stability problem associated with (4.3) can be thought of as a perturbation of that associated with the massive Thirring model (4.5). This will be seen to indeed be the case. As a consequence, much of the theory developed in [12, 13, 15, 16, 17] will be applicable. However, as shall be seen, some extensions of the theory are necessary in order to fully understand the problem.

We wish to linearize (4.3) about the wave given in Lemma 4.1. Thus, denoting the perturbation by $\mathbf{Y} = (u_1, u_1^*, u_2, u_2^*, a_3, a_3^*)^T \in \mathbb{C}^6$ and using the ansatz $\mathbf{Y}(x,t) = \mathbf{Y}(x)\mathrm{e}^{\mathrm{i}\lambda t}$ (so that the wave is unstable for $\mathrm{Im}\,\lambda < 0$), we obtain the eigenvalue problem

$$\begin{aligned} \mathrm{i}\partial_x u_1 + u_2 - (\lambda - \cos Q)u_1 - |\Phi|^2 u_2^* - \Phi a_3 &= 0,\\ \mathrm{i}\partial_x u_1^* - u_2^* - (\lambda + \cos Q)u_1^* + |\Phi|^2 u_2 + \Phi^* a_3^* &= 0,\\ -\mathrm{i}\partial_x u_2 + u_1 - |\Phi|^2 u_1^* - (\lambda - \cos Q)u_2 + \Phi^* a_3 &= 0,\\ -\mathrm{i}\partial_x u_2^* - u_1^* + |\Phi|^2 u_1 - (\lambda + \cos Q)u_2^* - \Phi a_3^* &= 0,\\ -(1 - 2\epsilon\cos Q)(\Phi^* u_1 - \Phi u_2) - (1 + \epsilon(\lambda - 2\cos Q))a_3 &= 0,\\ (1 - 2\epsilon\cos Q)(\Phi u_1^* - \Phi^* u_2^*) + (1 - \epsilon(\lambda + 2\cos Q))a_3^* &= 0. \end{aligned} \tag{4.6}$$

LEMMA 4.2. *If* λ *is an eigenvalue, then so is* $-\lambda$ *and* $\pm\lambda^*$*.*

Proof. The statement is a consequence of the fact that the system is Hamiltonian, which implies that eigenvalues appear as quadruples. A direct proof goes as follows: (4.6) is invariant under $(u_1, u_1^*, u_2, u_2^*, a_3, a_3^*) \mapsto (u_1^*, u_1, u_2^*, u_2, a_3^*, a_3)$ and $\lambda \to -\lambda$. Furthermore, these equations are invariant after taking the complex conjugate and then setting $\lambda \to -\lambda$. □

We investigate eigenvalues λ in bounded regions of $\mathbb{C}$ that are chosen independently of ϵ. For such λ, we can solve (4.6) directly for a_3 and a_3^*. Upon substituting the resulting expressions

$$a_3 = -\frac{1 - 2\epsilon\cos Q}{1 + \epsilon(\lambda - 2\cos Q)}(\Phi^* u_1 - \Phi u_2), \qquad a_3^* = -\frac{1 - 2\epsilon\cos Q}{1 - \epsilon(\lambda + 2\cos Q)}(\Phi u_1^* - \Phi^* u_2^*)$$

into the first four equations in (4.6), one gets the reduced eigenvalue problem

$$\begin{aligned} \partial_x u_1 &= \mathrm{i}\left[(\cos Q - \lambda + |\Phi|^2)u_1 + (1 - \Phi^2)u_2 - |\Phi|^2 u_2^*\right]\\ &\quad + \mathrm{i}\epsilon\lambda\left[-|\Phi|^2 u_1 + \Phi^2 u_2\right] + \mathrm{O}(\epsilon^2),\\ \partial_x u_1^* &= \mathrm{i}\left[-(\cos Q + \lambda + |\Phi|^2)u_1^* + |\Phi|^2 u_2 - (1 - (\Phi^*)^2)u_2^*\right]\\ &\quad + \mathrm{i}\epsilon\lambda\left[-|\Phi|^2 u_1^* + (\Phi^*)^2 u_2^*\right] + \mathrm{O}(\epsilon^2),\\ \partial_x u_2 &= \mathrm{i}\left[-(1 - (\Phi^*)^2)u_1 + |\Phi|^2 u_1^* - (\cos Q - \lambda + |\Phi|^2)u_2\right]\\ &\quad + \mathrm{i}\epsilon\lambda\left[-(\Phi^*)^2 u_1 + |\Phi|^2 u_2\right] + \mathrm{O}(\epsilon^2),\\ \partial_x u_2^* &= \mathrm{i}\left[-|\Phi|^2 u_1 + (1 - \Phi^2)u_1^* + (\cos Q + \lambda + |\Phi|^2)u_2^*\right]\\ &\quad + \mathrm{i}\epsilon\lambda\left[-\Phi^2 u_1^* + |\Phi|^2 u_2^*\right] + \mathrm{O}(\epsilon^2). \end{aligned} \tag{4.7}$$

Resetting $\mathbf{Y} = (u_1, u_1^*, u_2, u_2^*)^T \in \mathbb{C}^4$, we use the shortcut

$$\frac{\mathrm{d}\mathbf{Y}}{\mathrm{d}x} = M_0(\lambda, x)\mathbf{Y} + \epsilon M_\epsilon(\lambda, x)\mathbf{Y} \tag{4.8}$$

for the above system. Alternatively, we may write (4.7) in operator form as

$$[L_0 + \epsilon L_\epsilon]\mathbf{Y} = \lambda \mathbf{Y}, \tag{4.9}$$

where L_ϵ depends on λ.

4.3. The adjoint squared eigenfunctions of the massive Thirring model. The first step is to understand the eigenvalue problem for $\epsilon = 0$. Note that, when $\epsilon = 0$, the eigenvalue problem (4.9) is exactly that associated with the massive Thirring model. When $\epsilon = 0$, by exploiting integrability, we can conclude that the spectrum consists of an isolated eigenvalue at $\lambda = 0$ of geometric multiplicity two and algebraic multiplicity four, and continuous spectrum on the real axis with branch points at $\lambda = -1 \pm \cos Q$ and $\lambda = 1 \pm \cos Q$ (see [22, 23]). Since the spatial and rotational invariance associated with (4.3) persists under the perturbation, and since the problem is Hamiltonian, the eigenvalue at $\lambda = 0$ continues to have geometric multiplicity two and algebraic multiplicity four, even for $\epsilon \neq 0$. Any unstable eigenvalue must therefore bifurcate out of the continuous spectrum. The purpose of the rest of this section is to locate all bifurcating eigenvalues satisfying $|\lambda| \ll 1/\epsilon$.

The proof of the following lemma will be left to the interested reader, as it can easily be verified.

LEMMA 4.3. *Let*

$$\mathbf{Y}(\lambda, x) = [p_1(\lambda, x), p_2(\lambda, x), p_3(\lambda, x), p_4(\lambda, x)]^T$$

be a solution of (4.8) *for* $\epsilon = 0$. *Another solution is then given by*

$$\tilde{\mathbf{Y}}(\lambda, x) = [p_2^*(-\lambda, x), p_1^*(-\lambda, x), p_4^*(-\lambda, x), p_3^*(-\lambda, x)]^T.$$

Furthermore, if $\lambda \in \mathbb{R}$, *then*

$$\mathbf{Z}(\lambda, x) = [p_1(\lambda, x), -p_2(\lambda, x), -p_3(\lambda, x), p_4(\lambda, x)]^T$$

is a solution to the adjoint equation $\mathbf{Z}' = -M_0^*(\lambda, x)\mathbf{Z}$.

Edge bifurcations for the perturbed problem can be located by computing the Evans function for the unperturbed eigenvalue problem (4.8). Thus, as in section 3, we need to find solutions to the eigenvalue problem of the integrable massive Thirring model, i.e., to (4.8) or, equivalently, to (4.9). As shown in [22, 23], these solutions are given by the adjoint squared eigenfunctions of the scattering problem that is associated with the massive Thirring model. Hence we state only the result proved in [22, 23]. We emphasize, however, that it is again inverse scattering theory as outlined in section 2 that provides these solutions via the scattering problem, the Jost functions, and eventually the adjoint squared eigenfunctions.

PROPOSITION 4.4. *Let*

$$\Omega_f(k) = \cos Q + \sqrt{1 + k^2}, \qquad \Omega_s(k) = \cos Q - \sqrt{1 + k^2}.$$

When $\epsilon = 0$, *the eigenvalue problem* (4.9) *has two linearly independent solutions* $\Psi_s^A(k, x)$ *and* $\Psi_f^A(k, x)$ *such that*

$$L_0 \Psi_s^A(k, x) = \Omega_s(k)\Psi_s^A(k, x), \qquad L_0 \Psi_f^A(k, x) = -\Omega_f(k)\Psi_f^A(k, x).$$

Furthermore, these eigenfunctions have the asymptotics

$$\Psi_s^A(k,x)\mathrm{e}^{-\mathrm{i}kx} \longrightarrow \begin{cases} [1,0,-r(-k),0]^T, & x\to-\infty, \\ a_s^2(k)[1,0,-r(-k),0]^T, & x\to\infty, \end{cases}$$

$$\Psi_f^A(k,x)\mathrm{e}^{\mathrm{i}k^*x} \longrightarrow \begin{cases} [0,1,0,r(k)^*]^T, & x\to-\infty, \\ a_f^2(k)[0,1,0,r(k)^*]^T, & x\to\infty, \end{cases}$$

where the transmission coefficients

$$a_s(k) = \frac{r(-k)-\mathrm{e}^{\mathrm{i}Q}}{r(-k)-\mathrm{e}^{-\mathrm{i}Q}}\,\mathrm{e}^{-\mathrm{i}Q}, \qquad a_f(k) = \frac{r(k)^*+\mathrm{e}^{-\mathrm{i}Q}}{r(k)^*+\mathrm{e}^{\mathrm{i}Q}}\,\mathrm{e}^{\mathrm{i}Q},$$

with $r(k) = k+\sqrt{1+k^2}$, are such that $|a_s(k)| = |a_f(k)| = 1$ for real k.

Due to the simple relationship between (4.9) and (4.8), these eigenfunctions are also solutions to (4.8). Note that, as a consequence of Lemma 4.3, one also has eigenfunctions for $\lambda = -\Omega_f(k)$ and $\lambda = -\Omega_s(k)$. One should also note that the fact that there are two transmission coefficients for this problem is a reflection of the existence of two dispersion relationships $\lambda = \pm\Omega_s(k)$ and $\lambda = \pm\Omega_f(k)$.

4.4. Edge bifurcations for $Q \neq \pi/2$. In this section, it will be shown that the perturbed wave is linearly stable for $\cos Q \neq 0$. If $\cos Q = 0$, then additional technical difficulties are introduced which will be handled in the next section. Thus, throughout this section, we assume that $\cos Q \neq 0$. Let

$$M_\infty(\lambda) = \lim_{|x|\to\infty}(M_0(\lambda,x)+\epsilon M_\epsilon(\lambda,x)).$$

A routine calculation shows that the eigenvalues of $M_\infty(\lambda)$ are given by

$$\mu_s^\pm(\lambda) = \pm\sqrt{1-(\lambda-\cos Q)^2}, \qquad \mu_f^\pm(\lambda) = \pm\sqrt{1-(\lambda+\cos Q)^2}, \tag{4.10}$$

and the associated eigenvectors are

$$\mathbf{v}_s^\pm(\lambda) = [1,0,(\lambda-\cos Q)+\mathrm{i}\mu_s^\pm(\lambda),0]^T \tag{4.11}$$

$$\mathbf{v}_f^\pm(\lambda) = [0,1,0,-(\lambda+\cos Q)+\mathrm{i}\mu_f^\pm(\lambda)]^T. \tag{4.12}$$

The branch cuts for $\mu_{s,f}^\pm(\lambda)$ are taken so that the functions are analytic in the region $\operatorname{Im}\lambda \le 0$ except at the branch points $\lambda = \pm1\pm\cos Q$. For example, if $\operatorname{Re}\lambda < 0$, then

$$\arg(1-(\lambda-\cos Q)^2) \in (-\pi,\pi], \tag{4.13}$$

$$\arg(1-(\lambda+\cos Q)^2) \in (-\pi,\pi]. \tag{4.14}$$

Define solutions $\mathbf{Y}_{s,f}^+(\lambda,x)$ and $\mathbf{Y}_{s,f}^-(\lambda,x)$ such that

$$\lim_{x\to\infty}\mathbf{Y}_{s,f}^+(\lambda,x)\mathrm{e}^{-\mu_{s,f}^-(\lambda)x} = \mathbf{v}_{s,f}^-(\lambda), \qquad \lim_{x\to-\infty}\mathbf{Y}_{s,f}^-(\lambda,x)\mathrm{e}^{-\mu_{s,f}^+(\lambda)x} = \mathbf{v}_{s,f}^+(\lambda). \tag{4.15}$$

Note that, for $\operatorname{Im}\lambda < 0$, $\mathbf{Y}_{s,f}^-(\lambda,x)$ decays exponentially fast as $x\to-\infty$, while $\mathbf{Y}_{s,f}^+(\lambda,x)$ decays exponentially fast as $x\to\infty$. The Evans function is then given by

$$E(\lambda) = (\mathbf{Y}_s^-\wedge\mathbf{Y}_f^-\wedge\mathbf{Y}_s^+\wedge\mathbf{Y}_f^+)(\lambda,x). \tag{4.16}$$

To calculate the Evans function explicitly, we have to relate the information given in Proposition 4.4 to the solutions $\mathbf{Y}^{\pm}_{s,f}(\lambda, x)$. First consider the solution $\Psi^A_s(k, x)$. Inverting the relation $\lambda = \Omega_s(k)$ yields that Proposition 4.4 applies when $k = k_s(\lambda)$, where

$$k_s(\lambda) = \sqrt{(\lambda - \cos Q)^2 - 1}, \qquad \arg((\lambda - \cos Q)^2 - 1) \in (-2\pi, 0].$$

The restriction on the argument is so that the above expressions are consistent with (4.13). Note that, since $\Omega_s(k)$ is even in k, $\Psi^A_s(-k, x)$ is also a solution to the eigenvalue problem. An examination of the relations (4.10) and (4.11), as well as of the asymptotics given in (4.15), then reveals that

$$\mathbf{Y}^-_s(\lambda, x) = \Psi^A_s(k_s(\lambda), x), \qquad \mathbf{Y}^+_s(\lambda, x) = \frac{1}{a^2_s(-k_s(\lambda))}\,\Psi^A_s(-k_s(\lambda), x).$$

Now consider the solution $\Psi^A_f(k, x)$. As before, upon inverting $\lambda = \Omega_s(k)$, we see that Proposition 4.4 applies when $k = k_f(\lambda)$, where

$$k_f(\lambda) = \sqrt{(\lambda + \cos Q)^2 - 1}, \qquad \arg((\lambda + \cos Q)^2 - 1) \in (-2\pi, 0],$$

to make it consistent with (4.14). As above, $\Psi^A_f(-k, x)$ is also a solution to the eigenvalue problem since $\Omega_f(k)$ is even is k. Inspecting (4.10), (4.12), and (4.15), we obtain

$$\mathbf{Y}^-_f(\lambda, x) = \Psi^A_f(k_f(\lambda), x), \qquad \mathbf{Y}^+_f(\lambda, x) = \frac{1}{a^2_f(-k_f(\lambda))}\,\Psi^A_f(-k_f(\lambda), x).$$

We now have the following lemma, which follows immediately from the definition of the Evans function and the asymptotics of the squared eigenfunctions.

Lemma 4.5. *Assume that* $\cos Q \neq 0$. *The Evans function of* (4.8) *with* $\epsilon = 0$ *is given by*

$$E(\lambda) = -4k_s(\lambda)k_f(\lambda)\left[\frac{a_s(k_s(\lambda))}{a_s(-k_s(\lambda))}\right]^2\left[\frac{a_f(k_f(\lambda))}{a_f(-k_f(\lambda))}\right]^2.$$

Since $a_s(k_s(0)) = a_f(k_f(0)) = 0$, with each zero being simple, $\lambda = 0$ is a zero of multiplicity four for the Evans function. Furthermore, note that the Evans function has zeros at the branch points $\lambda = \pm 1 \pm \cos Q$. Assume that $\cos Q \neq 0$ so that the branch points of the Evans function do not coincide. As in [15], the Evans function can then be defined on a Riemann surface by setting

$$\gamma^2_s(\lambda) = 1 - (\lambda - \cos Q)^2 \qquad \text{or} \qquad \gamma^2_f(\lambda) = 1 - (\lambda + \cos Q)^2. \tag{4.17}$$

On each of these surfaces, the Evans function is analytic, and its zero at the branch point is simple. Therefore, the zero remains simple under perturbation. However, since the zero is simple and the system is Hamiltonian, the perturbed zero must either lie on the real axis or not correspond to an eigenvalue (i.e., it lies on the wrong sheet of the Riemann surface). In either case, it does not contribute to a linear instability.

4.5. Edge bifurcations for $Q = \pi/2$. The above conclusion of linear stability was predicated upon the fact that the branch points of the Evans function were separated. If $Q = \pi/2$, then the branch points, and hence the pair of simple zeros, coincide. It is then possible that complex zeros exist on the appropriate Riemann surface, which will necessarily lead to an unstable wave. We show in this section that the wave is indeed linearly unstable for $0 < Q - \pi/2 < 4\epsilon^2$, with the location of the unstable eigenvalues arising from the edge bifurcation being given by

$$\lambda = -1 + \epsilon^2 \left[2 - A - \mathrm{i}\sqrt{4A - A^2}\right], \tag{4.18}$$

where $Q - \pi/2 = A\epsilon^2$ with $0 < A < 4$. Otherwise, the wave is linearly stable, and only real eigenvalues arise from the edge bifurcation.

To calculate the Taylor expansion for the Evans function on the appropriate Riemann surface, the theory presented in [15, 17] must be slightly modified. First, rewrite the Evans function given in (4.16) as

$$E(\lambda) = ((\mathbf{Y}_s^- - \mathbf{Y}_s^+) \wedge (\mathbf{Y}_f^- - \mathbf{Y}_f^+) \wedge \mathbf{Y}_s^+ \wedge \mathbf{Y}_f^+)(\lambda, 0).$$

We wish to write the Evans function on the Riemann surface. The idea will be to look on the surface defined by $\gamma_s(\lambda)$, which is defined in (4.17). As a consequence, all expressions must be written in terms of this variable. Set

$$\nu^2 = 4\cos Q, \qquad \arg(\nu) \in (-\pi/2, \pi/2].$$

Expression (4.10) can now be rewritten as

$$\mu_s^\pm(\gamma_s) = \pm\gamma_s, \qquad \mu_f^\pm(\gamma_s) = \pm\sqrt{\gamma_s^2 + \nu^2\sqrt{1-\gamma_s^2} - \nu^4/4},$$

and expressions (4.11) and (4.12) have similar modifications. Note that when $\gamma_s = 0$ one has

$$\mu_f^\pm = \pm\nu\sqrt{1-\nu^2/4}, \qquad \mathbf{v}_f^\pm = \left(0, 1, 0, \sqrt{1-(\mu_f^\pm)^2} + \mathrm{i}\mu_f^\pm\right)^T,$$

and that the restriction on the argument of ν is consistent with that of (4.14).

We can now proceed to compute the Taylor expansion of the Evans function on the Riemann surface and, in particular, to get exact expressions for the coefficients in the expansion. The theory presented in [13, 15] will be heavily used and will not be rederived here. However, we will give the most important intermediate calculations. As an application of the gap lemma, the solutions $\mathbf{Y}_{s,f}^\pm(\lambda, x)$ can be written on the Riemann surface defined by $\gamma_s(\lambda)$. This will now be done implicitly, with the observation that $\lambda = -1$ corresponds to $\gamma_s = 0$. Let $\mathbf{u}_j(x)$, $j = 1, \ldots, 4$ be solutions to (4.8) with $(\epsilon, \nu) = (0, 0)$ such that $\mathbf{u}_1(x) = \mathbf{Y}_s^\pm(0, x)$, $\mathbf{u}_2(x) = \mathbf{Y}_f^\pm(0, x)$, and $\mathbf{u}_3(x)$ and $\mathbf{u}_4(x)$ are such that

$$(\mathbf{u}_1 \wedge \cdots \wedge \mathbf{u}_4)(x) = 1.$$

Let the adjoint solutions $\mathbf{u}_j^A(x)$ satisfy $\mathbf{u}_j(x) \cdot \mathbf{u}_k^A(x) = \delta_{jk}$. For $(\gamma_s, \nu) = (0, 0)$, the relevant eigenfunctions are given by

$$\tag{4.19} \begin{aligned} \mathbf{Y}_s^\pm(0, x) &= \begin{bmatrix} \tanh(2x)\tanh(x - \mathrm{i}\pi/4) \\ \mathrm{i}\,\mathrm{sech}(2x)\tanh(x + \mathrm{i}\pi/4) \\ -\tanh(2x)\tanh(x + \mathrm{i}\pi/4) \\ \mathrm{i}\,\mathrm{sech}(2x)\tanh(x - \mathrm{i}\pi/4) \end{bmatrix}, \\ \mathbf{Y}_f^\pm(0, x) &= \begin{bmatrix} \mathrm{i}\,\mathrm{sech}(2x)\tanh(x - \mathrm{i}\pi/4) \\ \tanh(2x)\tanh(x + \mathrm{i}\pi/4) \\ -\mathrm{i}\,\mathrm{sech}(2x)\tanh(x + \mathrm{i}\pi/4) \\ \tanh(2x)\tanh(x - \mathrm{i}\pi/4) \end{bmatrix}, \end{aligned}$$

while the relevant adjoint eigenfunctions $\mathbf{Z}_{s,f}(0,x)$ can be found by using the result of Lemma 4.3. It is important to note that the adjoint eigenfunctions are orthogonal to $\mathbf{Y}^{\pm}_{s,f}(0,x)$. Thus there exist constants c_{jk} such that

$$\mathbf{u}_3^A(x) = c_{11}\mathbf{Z}_s(0,x) + c_{12}\mathbf{Z}_f(0,x), \qquad \mathbf{u}_4^A(x) = c_{21}\mathbf{Z}_s(0,x) + c_{22}\mathbf{Z}_f(0,x).$$

Furthermore, the matrix $C = [c_{jk}] \in \mathbb{C}^{2\times 2}$ is invertible. As will be seen, the exact values for the entries c_{jk} are unimportant. The expressions in (4.19) will be implicitly used in the subsequent calculations.

When performing the perturbation expansion on the Riemann surface, the parameters that will be varied are γ_s, ν, and ϵ. It is important to keep in mind that $\gamma_f = \gamma_s$ when $\nu = 0$ so that in this case we will not distinguish between the two. For $j = 3, 4$, set

$$\alpha_j^{s,f} = \partial_{\gamma_s}\mathbf{v}^-_{s,f}\cdot\mathbf{u}_j^A(-\infty) - \partial_{\gamma_s}\mathbf{v}^+_{s,f}\cdot\mathbf{u}_j^A(+\infty).$$

Following [15], one has that

$$\partial_{\gamma_s}(\mathbf{Y}^-_{s,f} - \mathbf{Y}^+_{s,f})(0,0) = \sum_{j=3}^{4}\alpha_j^{s,f}\mathbf{u}_j(0) + \sum_{j=1}^{2}c^j_{s,f}\mathbf{u}_j(0)$$

on the Riemann surface. Using the fact that

$$\partial^2_{\gamma_s}E(0) = 2(\partial_{\gamma_s}(\mathbf{Y}^-_s - \mathbf{Y}^+_s)\wedge\partial_{\gamma_s}(\mathbf{Y}^-_f - \mathbf{Y}^+_f)\wedge\mathbf{Y}^+_s\wedge\mathbf{Y}^+_f)(0,0)$$

at $\gamma_s = 0$ and also utilizing the definition of the solutions $\mathbf{u}_j(0)$, we obtain

$$\partial^2_{\gamma_s}E(0) = 2\begin{vmatrix} \alpha_3^s & \alpha_4^s \\ \alpha_3^f & \alpha_4^f \end{vmatrix} = 8\det(C).$$

We must now get expressions for the lower-order derivatives in the Taylor expansion for $E(\gamma_s)$. Following [13], we have

$$\partial_\epsilon(\mathbf{Y}^-_{s,f} - \mathbf{Y}^+_{s,f})(0,0) = \sum_{j=3}^{4}\langle M_\epsilon\mathbf{Y}_{s,f},\mathbf{u}_j^A\rangle\mathbf{u}_j(0) + \sum_{j=1}^{2}c^j_{s,f}\mathbf{u}_j(0),$$

where $\langle\cdot,\cdot\rangle$ represents that standard L^2 inner product. Since

$$\partial^2_\epsilon E(0) = 2(\partial_\epsilon(\mathbf{Y}^-_s - \mathbf{Y}^+_s)\wedge\partial_\epsilon(\mathbf{Y}^-_f - \mathbf{Y}^+_f)\wedge\mathbf{Y}^+_s\wedge\mathbf{Y}^+_f)(0,0),$$

this then yields that

$$\partial^2_\epsilon E(0) = 2\begin{vmatrix} \langle M_\epsilon\mathbf{Y}_s,\mathbf{u}_3^A\rangle & \langle M_\epsilon\mathbf{Y}_s,\mathbf{u}_4^A\rangle \\ \langle M_\epsilon\mathbf{Y}_f,\mathbf{u}_3^A\rangle & \langle M_\epsilon\mathbf{Y}_f,\mathbf{u}_4^A\rangle \end{vmatrix} = -32\det(C).$$

Also, $\mathbf{Y}^-_s = \mathbf{Y}^+_s$ for all ν, while

$$\partial_\nu(\mathbf{Y}^-_f - \mathbf{Y}^+_f)(0,0) = \beta_3^f\mathbf{u}_3(0) + \beta_4^f\mathbf{u}_4(0) + c_f^1\mathbf{u}_1(0) + c_f^2\mathbf{u}_2(0),$$

where

$$\beta_j^f = \partial_\nu\mathbf{v}^-_f\cdot\mathbf{u}_j^A(-\infty) - \partial_\nu\mathbf{v}^+_f\cdot\mathbf{u}_j^A(+\infty).$$

This yields that $\partial_\nu^k E(0) = 0$ for all $k \geq 0$, while

$$\begin{aligned}
\partial^2_{\epsilon\gamma_s} E(0) &= (\partial_\epsilon(\mathbf{Y}_s^- - \mathbf{Y}_s^+) \wedge \partial_{\gamma_s}(\mathbf{Y}_f^- - \mathbf{Y}_f^+) \wedge \mathbf{Y}_s^+ \wedge \mathbf{Y}_f^+)(0,0) \\
&\quad + (\partial_{\gamma_s}(\mathbf{Y}_s^- - \mathbf{Y}_s^+) \wedge \partial_\epsilon(\mathbf{Y}_f^- - \mathbf{Y}_f^+) \wedge \mathbf{Y}_s^+ \wedge \mathbf{Y}_f^+)(0,0) \\
&= \begin{vmatrix} \langle M_\epsilon \mathbf{Y}_s, \mathbf{u}_3^A \rangle & \langle M_\epsilon \mathbf{Y}_s, \mathbf{u}_4^A \rangle \\ \alpha_3^f & \alpha_4^f \end{vmatrix} + \begin{vmatrix} \alpha_3^s & \alpha_4^s \\ \langle M_\epsilon \mathbf{Y}_f, \mathbf{u}_3^A \rangle & \langle M_\epsilon \mathbf{Y}_f, \mathbf{u}_4^A \rangle \end{vmatrix} \;=\; 0, \\
\partial^2_{\epsilon\nu} E(0) &= (\partial_\epsilon(\mathbf{Y}_s^- - \mathbf{Y}_s^+) \wedge \partial_\nu(\mathbf{Y}_f^- - \mathbf{Y}_f^+) \wedge \mathbf{Y}_s^+ \wedge \mathbf{Y}_f^+)(0,0) \\
&= \begin{vmatrix} \langle M_\epsilon \mathbf{Y}_s, \mathbf{u}_3^A \rangle & \langle M_\epsilon \mathbf{Y}_s, \mathbf{u}_4^A \rangle \\ \beta_3^f & \beta_4^f \end{vmatrix} \;=\; 0, \\
\partial^2_{\gamma_s\nu} E(0) &= (\partial_{\gamma_s}(\mathbf{Y}_s^- - \mathbf{Y}_s^+) \wedge \partial_\nu(\mathbf{Y}_f^- - \mathbf{Y}_f^+) \wedge \mathbf{Y}_s^+ \wedge \mathbf{Y}_f^+)(0,0) \\
&= \begin{vmatrix} \alpha_3^s & \alpha_4^s \\ \beta_3^f & \beta_4^f \end{vmatrix} \;=\; 4\det(C).
\end{aligned}$$

Thus, on the Riemann surface, the Evans function has the Taylor expansion

$$E(\gamma_s) = 4\det(C)\,(\gamma_s^2 + \nu\gamma_s - 4\epsilon^2),$$

and the zeros of the Evans function on the Riemann surface are given by

$$\gamma_s = \frac{1}{2}\left(-\nu \pm \sqrt{\nu^2 + 16\epsilon^2}\right).$$

Only zeros on the correct Riemann sheet, given by $\arg(\gamma_s) \in (-\pi/2, \pi/2]$, correspond to eigenvalues. Suppose that $\nu = 2\sqrt{\cos Q}$, i.e., $0 < \pi/2 - Q \ll 1$. The zeros are then real; furthermore, only one lies on the correct sheet. Upon inverting the relationship in (4.17), i.e., by setting

$$\lambda = \cos Q - \sqrt{1 - \gamma_s^2},$$

it is then seen that the real eigenvalue is, to lowest order, given by

$$\lambda = -1 + 2\cos Q + 2\epsilon^2 - \sqrt{\cos^2 Q + 4\epsilon^2 \cos Q}.$$

Next suppose that $0 < Q - \pi/2 \ll 1$, i.e., $\nu = 2\mathrm{i}\sqrt{-\cos Q}$. The zeros on the Riemann surface are then given by

$$\gamma_s = -\mathrm{i}\sqrt{-\cos Q} \pm \sqrt{4\epsilon^2 + \cos Q}.$$

The root lies on the correct sheet only for $0 < -\cos Q < 4\epsilon^2$, in which case it is given by

$$\gamma_s = -\mathrm{i}\sqrt{-\cos Q} + \sqrt{4\epsilon^2 + \cos Q}.$$

Upon setting $\cos Q = -A\epsilon^2$ for $A > 0$, i.e., $Q = \pi/2 + A\epsilon^2$, we see that the corresponding complex eigenvalue is, to lowest order, given by

$$\lambda = -1 + \epsilon^2\left[2 - A - \mathrm{i}\sqrt{4A - A^2}\right]. \tag{4.20}$$

Thus the wave is unstable.

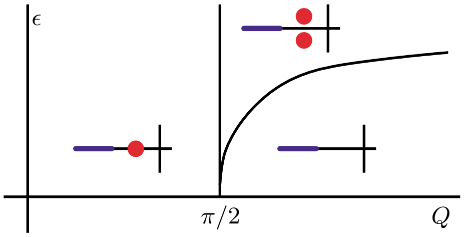


Fig. 4.1. *The insets show the eigenvalue structure near $\lambda = -1$ of the perturbed wave for Q close to $\pi/2$. The thick solid line in each inset denotes the essential spectrum, while bullets correspond to eigenvalues λ (i.e., to zeros of the Evans function on the correct Riemann sheet). Note that eigenvalues λ with $\operatorname{Im}\lambda < 0$ are unstable. The region of instability in the (Q, ϵ)-plane is bounded, to leading order, by the line $Q = \pi/2$ and the parabola $Q = \pi/2 + 4\epsilon^2$.*

In conclusion, the wave is linearly stable if $\cos Q < 0$ or if $\cos Q > -4\epsilon^2$. If $0 < -\cos Q < 4\epsilon^2$, then the wave is linearly unstable, and the location of the unstable eigenvalue is given to lowest order by (4.18). Note that, since the system is Hamiltonian, there is another unstable eigenvalue at $-\lambda^*$ with λ given by (4.20). This fact can be verified directly by computing the Taylor expansion of the Evans function about $\gamma_f = 0$. Our results near $Q = \pi/2$ are summarized in Figure 4.1.

Remark 4.6. Barashenkov, Pelinovsky, and Zemlyanaya [5] analyzed a different perturbation of the massive Thirring model and through the use of solvability conditions realized a different type of edge-bifurcation scenario as Q passed through $\pi/2$. (Compare Figure 4.1 with [5, Figure 1].) The results presented herein can be thought of as a theoretical justification for the calculations in [5].

Acknowledgments. We thank Alejandro Aceves for asking the question which led to this work and acknowledge helpful discussions with James Alexander in an early stage of this work.

REFERENCES

[1] M. Ablowitz and P. Clarkson, *Solitons, Nonlinear Evolution Equations, and Inverse Scattering*, London Math. Soc. Lecture Note Ser. 149, Cambridge University Press, Cambridge, UK, 1991.

[2] M. Ablowitz, D. Kaup, A. Newell, and H. Segur, *The inverse scattering transform—Fourier analysis for nonlinear problems*, Stud. Appl. Math., 53 (1974), pp. 249–315.

[3] N. Akhmediev and A. Ankiewicz, *Solitons: Nonlinear Pulses and Beams*, Chapman and Hall, London, 1997.

[4] J. Alexander, R. Gardner, and C. Jones, *A topological invariant arising in the stability of travelling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.

[5] I. Barashenkov, D. Pelinovsky, and E. Zemlyanaya, *Vibrations and oscillatory instabilities of gap solitons*, Phys. Rev. Lett., 80 (1998), pp. 5117–5120.

[6] J. Evans, *Nerve axon equations* I: *Linear approximations*, Indiana Univ. Math. J., 21 (1972), pp. 877–955.

[7] J. Evans, *Nerve axon equations* II: *Stability at rest*, Indiana Univ. Math. J., 22 (1972), pp. 75–90.

[8] J. Evans, *Nerve axon equations* III: *Stability of the nerve impulse*, Indiana Univ. Math. J., 22 (1972), pp. 577–594.

[9] J. Evans, *Nerve axon equations* IV: *The stable and unstable impulse*, Indiana Univ. Math. J., 24 (1975), pp. 1169–1190.

[10] R. Gardner and K. Zumbrun, *The gap lemma and geometric criteria for instability of viscous shock profiles*, Comm. Pure Appl. Math., 51 (1998), pp. 797–855.

[11] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer, New York, 1981.

[12] T. Kapitula, *Stability criterion for bright solitary waves of the perturbed cubic-quintic Schrödinger equation*, Phys. D, 116 (1998), pp. 95–120.

[13] T. Kapitula, *The Evans function and generalized Melnikov integrals*, SIAM J. Math. Anal., 30 (1998), pp. 273–297.

[14] T. Kapitula, *Stability of waves in perturbed Hamiltonian systems*, Phys. D, 156 (2001), pp. 186–200.

[15] T. Kapitula and J. Rubin, *Existence and stability of standing hole solutions to complex Ginzburg–Landau equations*, Nonlinearity, 13 (2000), pp. 77–112.

[16] T. Kapitula and B. Sandstede, *Instability mechanism for bright solitary-wave solutions to the cubic-quintic Ginzburg–Landau equation*, J. Opt. Soc. Amer. B Opt. Phys., 15 (1998), pp. 2757–2762.

[17] T. Kapitula and B. Sandstede, *Stability of bright solitary wave solutions to perturbed nonlinear Schrödinger equations*, Phys. D, 124 (1998), pp. 58–103.

[18] T. Kato, *Perturbation Theory for Linear Operators*, Springer, Berlin, 1980.

[19] D. J. Kaup, *Closure of the squared Zakharov–Shabat eigenstates*, J. Math. Anal. Appl., 54 (1976), pp. 849–864.

[20] D. J. Kaup, *A perturbation expansion for the Zakharov–Shabat inverse scattering transform*, SIAM J. Appl. Math., 31 (1976), pp. 121–133.

[21] D. J. Kaup, *Perturbation theory for solitons in optical fibers*, Phys. Rev. A (3), 42 (1990), pp. 5689–5694.

[22] D. J. Kaup and T. Lakoba, *The squared eigenfunctions of the massive Thirring model in laboratory coordinates*, J. Math. Phys., 37 (1996), pp. 308–323.

[23] D. J. Kaup and T. Lakoba, *Variational method: How it can create false instabilities*, J. Math. Phys., 37 (1996), pp. 3442–3462.

[24] D. J. Kaup and A. Newell, *Evolution equations, singular dispersion relations, and moving eigenvalues*, Adv. Math., 31 (1979), pp. 67–100.

[25] Y. Kivshar, D. Pelinovsky, T. Cretegny, and M. Peyrard, *Internal modes of solitary waves*, Phys. Rev. Lett., 80 (1998), pp. 5032–5035.

[26] Y. Kodama, M. Romagnoli, and S. Wabnitz, *Soliton stability and interactions in fibre lasers*, Elect. Lett., 28 (1992), pp. 1981–1983.

[27] Y. A. Li and K. Promislow, *The mechanism of the polarizational mode instability in birefringent fiber optics*, SIAM J. Math. Anal., 31 (2000), pp. 1351–1373.

[28] B. Malomed and R. Tasgal, *Vibration modes of a gap soliton in a nonlinear optical medium*, Phys. Rev. E (3), 49 (1994), pp. 5787–5796.

[29] B. Malomed and R. Tasgal, *The inverse scattering transform*, in Solitons, R. Bullough and P. Caudrey, eds., Springer, Berlin, 1980, pp. 177–242.

[30] R. Pego and M. Weinstein, *Eigenvalues, and instabilities of solitary waves*, Philos. Trans. Roy. Soc. London Ser. A, 340 (1992), pp. 47–94.

[31] D. Pelinovsky, Y. Kivshar, and V. Afanasjev, *Internal modes of envelope solitons*, Phys. D, 116 (1998), pp. 121–142.

[32] D. Pelinovsky and C. Sulem, *Bifurcations of new eigenvalues for the Benjamin–Ono equation*, J. Math. Phys., 39 (1998), pp. 6552–6572.

[33] D. Pelinovsky and C. Sulem, *Eigenfunctions and eigenvalues for a scalar Riemann–Hilbert problem associated to inverse scattering*, Comm. Math. Phys., 208 (2000), pp. 713–760.

[34] A. Rossi, C. Conti, and S. Trillo, *Stability, multistability, and wobbling of optical gap solitons*, Phys. Rev. Lett., 81 (1998), pp. 85–88.

[35] W. van Saarloos and P. Hohenberg, *Fronts, pulses, sources, and sinks in the generalized complex Ginzburg–Landau equation*, Phys. D, 56 (1992), pp. 303–367.

[36] B. Sandstede, *Stability of travelling waves*, in Handbook of Dynamical Systems: Towards Applications, B. Fiedler, ed., Elsevier, Amsterdam, in press.

[37] S. Trillo, Personal communication, University of Ferrara, Ferrara, Italy, 1999.

[38] S. Trillo, *Resonance Thirring solitons in type* II *second-harmonic generation*, Opt. Lett., 21 (1996), pp. 1732–1734.

[39] M. I. Weinstein, *Modulational stability of ground states of nonlinear Schrödinger equations*, SIAM J. Math. Anal., 16 (1985), pp. 472–491.

[40] J. Yang and D. J. Kaup, *Stability and evolution of solitary waves in perturbed generalized nonlinear Schrödinger equations*, SIAM J. Appl. Math., 60 (2000), pp. 967–989.

SIAM J. MATH. ANAL.
Vol. 33, No. 5, pp. 1144–1165

GEOMETRIC STABILITY SWITCH CRITERIA IN DELAY DIFFERENTIAL SYSTEMS WITH DELAY DEPENDENT PARAMETERS*

EDOARDO BERETTA† AND YANG KUANG‡

Abstract. In most applications of delay differential equations in population dynamics, the need of incorporation of time delays is often the result of the existence of some stage structure. Since the through-stage survival rate is often a function of time delays, it is easy to conceive that these models may involve some delay dependent parameters. The presence of such parameters often greatly complicates the task of an analytical study of such models. The main objective of this paper is to provide practical guidelines that combine graphical information with analytical work to effectively study the local stability of some models involving delay dependent parameters. Specifically, we shall show that the stability of a given steady state is simply determined by the graphs of some functions of τ which can be expressed explicitly and thus can be easily depicted by Maple and other popular software. In fact, for most application problems, we need only look at one such function and locate its zeros. This function often has only two zeros, providing thresholds for stability switches. The common scenario is that as time delay increases, stability changes from stable to unstable to stable, implying that a large delay can be stabilizing. This scenario often contradicts the one provided by similar models with only delay independent parameters.

Key words. delay differential equations, stability switch, characteristic equations, stage structure, population models

AMS subject classifications. 34K18, 34K20, 92D25

PII. S0036141000376086

1. Introduction.

Due to the fact that actions and reactions take time to take effect in real-life problems, one often introduces time delays in the variables being modeled. This often yields delay differential and delay difference models [11], [21], [19]. Some of these models have delay dependent parameters (for example, [1], [2], [3], [4], [9], [10], [22]), while most of them contain only parameters that are independent of time delays.

In most applications of delay differential equations in population dynamics, the need of incorporation of a time delay is often the result of the existence of some stage structure [1], [3], [10], [11], [12], [15]. Indeed, just about every population goes through some distinct life stages [23], [18]. Since the through-stage survival rate is often a function of a time delay, it is thus easy to conceive that these models will inevitably involve some delay dependent parameters.

In view of the fact that it is often difficult to analytically study models with delay dependent parameters even if only a single discrete delay is present, it is natural to resort to the help of computer programs. The main objective of this paper is to provide practical guidelines that combine graphical information with analytical work to effectively study the local stability of models involving delay dependent parameters. To apply our results, one need only perform some routine computation (using our analytical criteria) and generate some simple graphs which can be easily produced by popular software such as Maple. The results also can be readily confirmed by some

*Received by the editors July 28, 2000; accepted for publication (in revised form) September 4, 2001; published electronically February 14, 2002.

http://www.siam.org/journals/sima/33-5/37608.html

†Istituto di Biomatematica, Universita di Urbino, I-61029 Urbino, Italy (e.beretta@mat.uniurb.it).

‡Department of Mathematics, Arizona State University, Tempe, AZ 85287 (kuang@asu.edu). This author's work was partially supported by NSF grant DMS-0077790.

selective simulations using the freely available and user friendly software XPP. No other programming skill is required.

Specifically, we shall show that the stability of a given steady state is simply determined by the graphs of some functions of τ which can be expressed explicitly and thus can be easily depicted by Maple and other popular software. In fact, for most application problems, we need only look at one such function and locate its zeros. This function often has only two zeros, providing thresholds for stability switches. The common scenario is that as time delay increases, stability changes from stable to unstable to stable. We hope this work will show that it is important and possible to systematically study local stability aspects of some models with delay dependent parameters.

In the next section, we present a general geometric criterion that, theoretically speaking, can be applied to models with many delays, or even distributed delays [5], [7]. This is followed by a section dealing with the simple case of a first order characteristic equation, providing more user friendly geometric and analytic criteria for stability switches. In section 4, we accomplish the same for the second order case. The analytical criteria provided for the first and second order cases can be used to obtain some insightful analytical statements and can be helpful for conducting simulations. Examples are provided for both first and second order cases to illustrate the applications of our criteria. A discussion section concludes the paper.

2. A general geometric criterion.

In this section we study the occurrence of any possible stability switching *resulting from the increase of value of the time delay* τ for the general characteristic equation

$$D(\lambda,\tau)=0. \tag{2.1}$$

Here

$$D(\lambda,\tau)=P_n(\lambda,\tau)+Q_m(\lambda,\tau)e^{-\lambda\tau} \tag{2.2}$$

and

$$P_n(\lambda,\tau)=\sum_{k=0}^{n}p_k(\tau)\lambda^k, \quad Q_m(\lambda,\tau)=\sum_{k=0}^{m}q_k(\tau)\lambda^k. \tag{2.3}$$

In (2.3), $n,m\in\mathbf{N}_0$, $n>m$, and $p_k(\cdot),q_k(\cdot):\mathbf{R}_{+0}\to\mathbf{R}$ are continuous and differentiable functions of τ such that

$$P_n(0,\tau)+Q_m(0,\tau)=p_0(\tau)+q_0(\tau)\neq 0 \quad \forall\tau\in\mathbf{R}_{+0}, \tag{2.4}$$

i.e., $\lambda=0$ is not a characteristic root of (2.1).

In the following "—" denotes complex and conjugate. $P_n(\lambda,\tau)$, $Q_m(\lambda,\tau)$ are analytic functions in λ and differentiable in τ for which we assume ([21, p. 83]; see also [8] and [6]) the following:

(i) If $\lambda=i\omega,\omega\in\mathbf{R}$, then $P_n(i\omega,\tau)+Q_m(i\omega,\tau)\neq 0,\tau\in\mathbf{R}$;

(ii) $\limsup\{|Q_m(\lambda,\tau)/P_n(\lambda,\tau)|:|\lambda|\to\infty,\mathrm{Re}\ \lambda\geq 0\}<1$ for any τ;

(iii) $F(\omega,\tau):=|P_n(i\omega,\tau)|^2-|Q_m(i\omega,\tau)|^2$ for each τ has at most a finite number of real zeros.

(iv) Each positive root $\omega(\tau)$ of $F(\omega,\tau)=0$ is continuous and differentiable in τ whenever it exists.

Assumption (i) implies that $P_n(\lambda, \tau)$ and $Q_m(\lambda, \tau)$ have no common imaginary roots. This is needed to ensure that threshold time delays can be explicitly expressed (see the comment after (2.10)). Assumption (ii) is needed to ensure that there are no roots bifurcating from infinity. Assumption (iii) is needed to ensure that there are only finite "gates" for roots to cross the imaginary axis for any given τ. Assumption (iv) is needed to compute the derivative of the imaginary roots with respect to τ. We remark also that, since P_n, Q_m have real coefficients, then

$$\overline{P_n(-i\omega, \tau)} = P_n(i\omega, \tau), \qquad \overline{Q_m(-i\omega, \tau)} = Q_m(i\omega, \tau) \tag{2.5}$$

for each τ and any real ω, thus ensuring that if $\lambda = i\omega$ for some real ω is a characteristic root of (2.1), then also $\lambda = -i\omega$ is a characteristic root. In the following, we will drop the indices n, m from P_n, Q_m. Furthermore, we will denote by P_R, Q_R the real parts of P and Q, respectively, and by P_I, Q_I the imaginary parts of P and Q, respectively. Hence, we can write

$$P(\lambda, \tau) = P_R(\lambda, \tau) + iP_I(\lambda, \tau), \quad Q(\lambda, \tau) = Q_R(\lambda, \tau) + iQ_I(\lambda, \tau), \tag{2.6}$$

where P_R, P_I, Q_R, Q_I are real functions. In the following we will use this nomenclature for derivatives. The total derivative, say of $P(\lambda, \tau)$, with respect to τ will be denoted by

$$D_\tau P(\lambda, \tau) := P'_\lambda(\lambda, \tau)\frac{d\lambda}{d\tau} + P'_\tau(\lambda, \tau), \tag{2.7}$$

where $P'_\lambda(\lambda, \tau) := \partial_\lambda P(\lambda, \tau)$, $P'_\tau(\lambda, \tau) := \partial_\tau P(\lambda, \tau)$ are the partial derivatives with respect to λ, τ, respectively. Of course, after the partial derivation of $P(\lambda, \tau)$ with respect to λ, its real part can be separated from its imaginary one. We then have

$$P'_\lambda(\lambda, \tau) = P'_{R_\lambda}(\lambda, \tau) + iP'_{I_\lambda}(\lambda, \tau) \tag{2.8}$$

and the same nomenclature applies for derivatives of $Q(\lambda, \tau)$. Similarly, $F'_\omega(\omega, \tau) = \partial_\omega F(\omega, \tau)$ denotes the partial derivative of $F(\omega, \tau)$ with respect to ω, and so on. This stated, let us consider the general problem. Since for increasing τ the imaginary axis cannot be crossed by $\lambda(\tau) = 0$ for some $\tau > 0$ (see (2.4)), we look for the occurrence of a pair of simple and conjugate imaginary roots $\lambda = \pm i\omega(\tau)$, $\omega(\tau)$ real and positive, which crosses the imaginary axis at some positive τ value, say τ^*. Because of (2.5), without loss of generality, we can consider just $\lambda = i\omega(\tau)$, $\omega(\tau) > 0$, and the possibility that it is a root of the characteristic equation (2.1). Then $\omega(\tau)$ must satisfy the following:

$$\begin{cases} Q_I(i\omega, \tau)\sin\omega\tau + Q_R(i\omega, \tau)\cos\omega\tau = -P_R(i\omega, \tau), \\ -Q_R(i\omega, \tau)\sin\omega\tau + Q_I(i\omega, \tau)\cos\omega\tau = -P_I(i\omega, \tau), \end{cases} \tag{2.9}$$

which gives

$$\begin{cases} \sin\omega\tau = \dfrac{-P_R(i\omega, \tau)Q_I(i\omega, \tau) + P_I(i\omega, \tau)Q_R(i\omega, \tau)}{|Q(i\omega, \tau)|^2}, \\ \cos\omega\tau = -\dfrac{P_R(i\omega, \tau)Q_R(i\omega, \tau) + P_I(i\omega, \tau)Q_I(i\omega, \tau)}{|Q(i\omega, \tau)|^2}, \end{cases} \tag{2.10}$$

where $|Q(i\omega, \tau)|^2 \neq 0$ because of assumption (i) (since $D(i\omega, \tau) = Q(i\omega, \tau) = 0$ together imply $P(i\omega, \tau) = 0$).

On the other hand, since (2.10) can be written as

$$\sin\omega\tau = \operatorname{Im}\left(\frac{P(i\omega,\tau)}{Q(i\omega,\tau)}\right), \qquad \cos\omega\tau = -\operatorname{Re}\left(\frac{P(i\omega,\tau)}{Q(i\omega,\tau)}\right),$$

if ω satisfies (2.10), then $\omega(\tau)$ must satisfy that

$$|P(i\omega,\tau)|^2 = |Q(i\omega,\tau)|^2, \tag{2.11}$$

i.e., $\omega(\tau)$ must be a (positive) root of

$$F(\omega,\tau) := |P(i\omega,\tau)|^2 - |Q(i\omega,\tau)|^2. \tag{2.12}$$

Assume that $I \subseteq \mathbf{R}_{+0}$ is the set where $\omega(\tau)$ is a positive root of (2.12) and for $\tau \notin I$, $\omega(\tau)$ is not definite. Then for all τ in I, $\omega(\tau)$ satisfies that

$$F(\omega,\tau) = 0. \tag{2.13}$$

Hence, differentiating (2.13) with respect to τ we get

$$F'_\omega(\omega,\tau)\omega' + F'_\tau(\omega,\tau) = 0, \quad \tau \in I, \tag{2.14}$$

where

$$\begin{cases} F'_\omega = 2[(P'_{R_\omega}P_R + P'_{I_\omega}P_I) - (Q'_{R_\omega}Q_R + Q'_{I_\omega}Q_I)], \\ F'_\tau = 2[(P'_{R_\tau}P_R + P'_{I_\tau}P_I) - (Q'_{R_\tau}Q_R + Q'_{I_\tau}Q_I)]. \end{cases} \tag{2.15}$$

Now it is important to notice that if $\tau \notin I$, then there are no positive $\omega(\tau)$ solutions of (2.13), and we cannot have stability switches. Furthermore, for any $\tau \in I$ where $\omega(\tau)$ is a positive solution of (2.13), we can define the angle $\theta(\tau) \in [0, 2\pi]$ as the solution of (2.10):

$$\begin{cases} \sin\theta(\tau) = \dfrac{-P_R(i\omega,\tau)Q_I(i\omega,\tau) + P_I(i\omega,\tau)Q_R(i\omega,\tau)}{|Q(i\omega,\tau)|^2}, \\ \cos\theta(\tau) = -\dfrac{P_R(i\omega,\tau)Q_R(i\omega,\tau) + P_I(i\omega,\tau)Q_I(i\omega,\tau)}{|Q(i\omega,\tau)|^2}, \end{cases} \tag{2.16}$$

and the relation between the arguments "$\theta(\tau)$" in (2.16) and "$\omega(\tau)\tau$" in (2.10) for $\tau \in I$ must be

$$\omega(\tau)\tau = \theta(\tau) + n2\pi, \quad n \in \mathbf{N}_0. \tag{2.17}$$

Hence, we can define the maps $\tau_n : I \to \mathbf{R}_{+0}$ given by

$$\tau_n(\tau) := \frac{\theta(\tau) + n2\pi}{\omega(\tau)}, \quad n \in \mathbf{N}_0, \quad \tau \in I, \tag{2.18}$$

where $\omega(\tau)$ is a positive root of (2.13). Let us introduce the functions $I \to \mathbf{R}$,

$$S_n(\tau) := \tau - \tau_n(\tau), \quad \tau \in I, \quad n \in \mathbf{N}_0, \tag{2.19}$$

that are continuous and differentiable in τ as shown in the following lemma.

LEMMA 2.1. *Assume that $\omega(\tau)$ is a positive real root of* (2.13) *defined for $\tau \in I$, which is continuous and differentiable. Assume further that* (i) *holds true. Then the functions $S_n(\tau), n \in \mathbf{N}_0$, are continuous and differentiable on I.*

Proof. Assume that $\theta(\tau)$ is, for example, a monotone increasing function in a neighborhood $I_\delta(\tau')$ of $\tau' \in I$, where $\theta(\tau') = 2\pi$. Since $\theta(\tau)$ must belong to $[0, 2\pi]$ at τ', $\theta(\tau)$ may have a jump of height 2π down to 0 at τ'. This will give rise to the first kind of discontinuity for $\tau_n(\tau)$ and $S_n(\tau)$ with a jump of height $2\pi/\omega(\tau')$ at τ'. Without such a discontinuity, Remark 4.1 (see below) implies that $\theta(\tau)$ is continuous and differentiable on I and so are $\tau_n(\tau)$ and $S_n(\tau)$. Thus it is enough to prove that $\theta(\tau) \neq 0, 2\pi$ on I, and hence $\theta(\tau) \in (0, 2\pi)$ on I. Assumption (i) implies that either

(a) $P_R(i\omega, \tau) + Q_R(i\omega, \tau) \neq 0$, or

(b) $P_I(i\omega, \tau) + Q_I(i\omega, \tau) \neq 0$.

Assume first that (a) holds true. Then either

(a1) $P_R(i\omega, \tau) \neq 0$, or

(a2) $Q_R(i\omega, \tau) \neq 0$.

Assume now that (a1) is true. If $\theta(\tau) = 0, 2\pi$, then $\sin\theta(\tau) = 0$ and $\cos\theta(\tau) = 1$. From the first of (4.16), we have

$$Q_I = P_I Q_R / P_R.$$

Substituting this into the second equation of (4.16) yields (since ω is a root of $F(\omega, \tau) = 0$)

$$\cos\theta(\tau) = -Q_R/P_R.$$

Hence $\cos\theta(\tau) = 1$ implies that $P_R + Q_R = 0$, contradicting (a). The proof for case (a2) is similar and so is the proof for (b). This proves the lemma. □

We can also prove the following theorem.

THEOREM 2.2. *Assume that $\omega(\tau)$ is a positive real root of* (2.13) *defined for $\tau \in I$, $I \subseteq \mathbf{R}_{+0}$, and at some $\tau^* \in I$,*

$$S_n(\tau^*) = 0 \quad \textit{for some } n \in \mathbf{N}_0. \tag{2.20}$$

Then a pair of simple conjugate pure imaginary roots $\lambda_+(\tau^*) = i\omega(\tau^*), \lambda_-(\tau^*) = -i\omega(\tau^*)$ of (2.1) *exists at $\tau = \tau^*$ which crosses the imaginary axis from left to right if $\delta(\tau^*) > 0$ and crosses the imaginary axis from right to left if $\delta(\tau^*) < 0$, where*

$$\delta(\tau^*) = \operatorname{sign}\left\{\left.\frac{d\mathrm{Re}\ \lambda}{d\tau}\right|_{\lambda = i\omega(\tau^*)}\right\} = \operatorname{sign}\{F'_\omega(\omega(\tau^*), \tau^*)\}\operatorname{sign}\left\{\left.\frac{dS_n(\tau)}{d\tau}\right|_{\tau = \tau^*}\right\}. \tag{2.21}$$

Proof. The existence part of the theorem follows from the requirement (2.17) which ensures that if and only if $\tau^* \in I$ is a zero of $S_n(\tau)$ for some $n \in \mathbf{N}_0$, $\lambda = \pm i\omega(\tau^*)$ together with $\omega(\tau^*) > 0$, a solution of (2.13), are characteristic roots of (2.1).

To prove the geometric criterion (2.21) we remark that

$$\operatorname{sign}\left\{\frac{d\mathrm{Re}\lambda}{d\tau}\right\} = \operatorname{sign}\left\{\mathrm{Re}\left(\frac{d\lambda}{d\tau}\right)^{-1}\right\}.$$

Then differentiating (2.1) with respect to τ we obtain that

$$\begin{aligned}&\left(\frac{d\lambda}{d\tau}\right)[P'_\lambda(\lambda, \tau) + (Q'_\lambda(\lambda, \tau) - \tau Q(\lambda, \tau))e^{-\lambda\tau}]\\ &= \lambda Q(\lambda, \tau)e^{-\lambda\tau} - [P'_\tau(\lambda, \tau) + Q'_\tau(\lambda, \tau)e^{-\lambda\tau}].\end{aligned} \tag{2.22}$$

From (2.2), we have

$$e^{\lambda\tau} = -\frac{Q(\lambda,\tau)}{P(\lambda,\tau)}. \tag{2.23}$$

Hence, we obtain

$$\left(\frac{d\lambda}{d\tau}\right)^{-1} = \left(-\frac{P'_\lambda(\lambda,\tau)}{P(\lambda,\tau)} + \frac{Q'_\lambda(\lambda,\tau)}{Q(\lambda,\tau)} - \tau\right) \Big/ \left(\lambda + \frac{P'_\tau(\lambda,\tau)}{P(\lambda,\tau)} - \frac{Q'_\tau(\lambda,\tau)}{Q(\lambda,\tau)}\right), \tag{2.24}$$

where $P(i\omega,\tau)$, $Q(i\omega,\tau) \neq 0$ due to assumption (i). Assume that $\lambda = i\omega(\tau)$, where $\omega(\tau) > 0$ is a root of (2.13). Then from (2.24) we obtain

$$\begin{aligned}&\operatorname{sign}\left\{\left.\frac{d\mathrm{Re}\,\lambda}{d\tau}\right|_{\lambda=i\omega}\right\}\\ &\quad = \operatorname{sign}\,\mathrm{Re}\left\{\frac{-P'_\lambda(i\omega,\tau)\overline{P(i\omega,\tau)} + Q'_\lambda(i\omega,\tau)\overline{Q(i\omega,\tau)} - \tau|P(i\omega,\tau)|^2}{P'_\tau(i\omega,\tau)\overline{P(i\omega,\tau)} - Q'_\tau(i\omega,\tau)\overline{Q(i\omega,\tau)} + i\omega|P(i\omega,\tau)|^2}\right\}.\end{aligned} \tag{2.25}$$

Now we remark that

$$iP'_\lambda(i\omega,\tau) = P'_\omega(i\omega,\tau), \qquad iQ'_\lambda(i\omega,\tau) = Q'_\omega(i\omega,\tau). \tag{2.26}$$

Hence, in (2.25) we have

$$\begin{aligned}&-P'_\lambda(i\omega,\tau)\overline{P(i\omega,\tau)} + Q'_\lambda(i\omega,\tau)\overline{Q(i\omega,\tau)}\\ &\quad = i[(P'_{R_\omega}P_R + P'_{I_\omega}P_I) - (Q'_{R_\omega}Q_R + Q'_{I_\omega}Q_I)]\\ &\qquad -[(P'_{I_\omega}P_R - P_IP'_{R_\omega}) - (Q'_{I_\omega}Q_R - Q_IQ'_{R_\omega})],\end{aligned} \tag{2.27}$$

which due to (2.15) becomes

$$\begin{aligned}&-P'_\lambda(i\omega,\tau)\overline{P(i\omega,\tau)} + Q'_\lambda(i\omega,\tau)\overline{Q(i\omega,\tau)}\\ &\quad = i\frac{F'_\omega(\omega,\tau)}{2} - [(P_RP'_{I_\omega} - P_IP'_{R_\omega}) - (Q_RQ'_{I_\omega} - Q_IQ'_{R_\omega})].\end{aligned} \tag{2.28}$$

Similarly, in (2.25) we have

$$\begin{aligned}&P'_\tau(i\omega,\tau)\overline{P(i\omega,\tau)} - Q'_\tau(i\omega,\tau)\overline{Q(i\omega,\tau)}\\ &\quad = \frac{1}{2}F'_\tau(\omega,\tau) + i[(P_RP'_{I_\tau} - P_IP'_{R_\tau}) - (Q_RQ'_{I_\tau} - Q_IQ'_{R_\tau})].\end{aligned} \tag{2.29}$$

Furthermore, remember that from (2.14)

$$F'_\tau(\omega,\tau) = -F'_\omega(\omega,\tau)\omega'. \tag{2.30}$$

Hence, from (2.28)–(2.30) in (2.25) we obtain

$$\operatorname{sign}\left\{\left.\frac{d\mathrm{Re}\,\lambda}{d\tau}\right|_{\lambda=i\omega}\right\} = \operatorname{sign}\,\mathrm{Re}\left\{\frac{-2\{U + \tau|P(i\omega,\tau)|^2\} + iF'_\omega(\omega,\tau)}{F'_\tau(\omega,\tau) + i2\{V + \omega|P(i\omega,\tau)|^2\}}\right\},$$

where

$$U := (P_RP'_{I_\omega} - P_IP'_{R_\omega}) - (Q_RQ'_{I_\omega} - Q_IQ'_{R_\omega}), \quad V := (P_RP'_{I_\tau} - P_IP'_{R_\tau}) - (Q_RQ'_{I_\tau} - Q_IQ'_{R_\tau}).$$

Simple computation yields

$$\text{sign}\left\{\left.\frac{d\text{Re }\lambda}{d\tau}\right|_{\lambda=i\omega}\right\} = \text{sign }\{F'_\omega(\omega,\tau)\}\text{ sign }\left\{\tau\omega' + \omega + \frac{U\omega'+V}{|P(i\omega,\tau)|^2}\right\}. \tag{2.31}$$

Now observe that if $S_n(\tau^*) = 0$, then $S'_n(\tau^*) = (\omega(\tau^*) + \tau^*\omega'(\tau^*) - \theta'(\tau^*))/\omega(\tau^*)$, which gives

$$\text{sign }\{S'_n(\tau^*)\} = \text{sign }\{\omega(\tau^*) + \tau^*\omega'(\tau^*) - \theta'(\tau^*)\}, \tag{2.32}$$

where $\theta'(\tau^*)$ can be computed with the help of (2.16). Let

$$\begin{cases} \psi(\tau) = -P_R(i\omega,\tau)Q_I(i\omega,\tau) + P_I(i\omega,\tau)Q_R(i\omega,\tau), \\ \varphi(\tau) = P_R(i\omega,\tau)Q_R(i\omega,\tau) + P_I(i\omega,\tau)Q_I(i\omega,\tau); \end{cases} \tag{2.33}$$

then for all $\tau \in I$, $\theta'(\tau)$ is defined as (see Remark 2.1)

$$\theta'(\tau) = \frac{\psi(\tau)\varphi'(\tau) - \psi'(\tau)\varphi(\tau)}{|P(i\omega,\tau)|^4}, \tag{2.34}$$

where

$$\begin{cases} \varphi'(\tau) = (P'_{R_\omega}Q_R + P_RQ'_{R_\omega} + P'_{I_\omega}Q_I + P_IQ'_{I_\omega})\omega' \\ \quad +(P'_{R_\tau}Q_R + P_RQ'_{R_\tau} + P'_{I_\tau}Q_I + P_IQ'_{I_\tau}), \\ \psi'(\tau) = (-P'_{R_\omega}Q_I - P_RQ'_{I_\omega} + Q'_{R_\omega}P_I + Q_RP'_{I_\omega})\omega' \\ \quad +(-P'_{R_\tau}Q_I - P_RQ'_{I_\tau} + Q'_{R_\tau}P_I + Q_RP'_{I_\tau}). \end{cases} \tag{2.35}$$

Hence, from (2.34), (2.35) we have

$$\begin{aligned} |P(i\omega,\tau)|^4\theta'(\tau) = \omega'\{&(P'_{R_\omega}Q_R + P_RQ'_{R_\omega} + P'_{I_\omega}Q_I + P_IQ'_{I_\omega})(-P_RQ_I + Q_RP_I) \\ &(-P'_{R_\omega}Q_I - P_RQ'_{I_\omega} + Q'_{R_\omega}P_I + Q_RP'_{I_\omega})(P_RQ_R + P_IQ_I)\} \\ &+(P'_{R_\tau}Q_R + P_RQ'_{R_\tau} + P'_{I_\tau}Q_I + P_IQ'_{I_\tau})(-P_RQ_I + Q_RP_I) \\ &-(-P'_{R_\tau}Q_I - P_RQ'_{I_\tau} + Q'_{R_\tau}P_I + Q_RP'_{I_\tau})(P_RQ_R + P_IQ_I) \\ &\equiv \omega'A + B. \end{aligned} \tag{2.36}$$

It can be shown that

$$\begin{aligned} A &= -(P_R^2 + P_I^2)Q'_{R_\omega}Q_I - (Q_R^2 + Q_I^2)P_RP'_{I_\omega} + P'_{R_\omega}P_I(Q_R^2 + Q_I^2) \\ &\quad +(P_R^2 + P_I^2)Q_RQ'_\omega \\ &= (P'_{R_\omega}P_I - P_RP'_{I_\omega})|P(i\omega,\tau)|^2 - (Q'_{R_\omega}Q_I - Q_RQ'_{I_\omega})|P(i\omega,\tau)|^2 \end{aligned} \tag{2.37}$$

and

$$\begin{aligned} B &= -(P_R^2 + P_I^2)(Q'_{R_\tau}Q_I - Q_RQ'_{I_\tau}) + (Q_R^2 + Q_I^2)(P'_{R_\tau}P_I - P_RP'_{I_\tau}) \\ &= |P(i\omega,\tau)|^2(Q_RQ'_{I_\tau} - Q'_{R_\tau}Q_I) - |P(i\omega,\tau)|^2(P_RP'_{I_\tau} - P'_{R_\tau}P_I). \end{aligned} \tag{2.38}$$

Hence, from (2.36)–(2.38) we obtain

$$\begin{aligned} \theta'(\tau) &= -\frac{\omega'[(P'_{I_\omega}P_R - P'_{R_\omega}P_I) - (Q_RQ'_{I_\omega} - Q'_{R_\omega}Q_I)]}{|P(i\omega,\tau)|^2} \\ &\quad -\frac{[(P_RP'_{I_\tau} - P'_{R_\tau}P_I) - (Q_RQ'_{I_\tau} - Q'_{R_\tau}Q_I)]}{|P(i\omega,\tau)|^2} = -\frac{U\omega'+V}{|P(i\omega,\tau)|^2}. \end{aligned} \tag{2.39}$$

Therefore, if we substitute (2.39) evaluated at τ^* in (2.32), and we compare the result with (2.31), we find that

$$\operatorname{sign}\left\{\frac{d\mathrm{Re}\ \lambda}{d\tau}\bigg|_{\lambda=i\omega(\tau^*)}\right\} = \operatorname{sign}\ \{F'_\omega(\omega,\tau^*)\}\operatorname{sign}\ \{S'_n(\tau^*)\},$$

thus completing the proof. □

Remark 2.1. Assume that $\theta(\tau)\in(0,2\pi), \tau\in I$, where $\theta(\tau)$ is defined by (2.16). According to (2.35), we can rewrite (2.16) as

$$\sin\theta(\tau) = \frac{\psi(\tau)}{|Q(i\omega,\tau)|^2}, \qquad \cos\theta(\tau) = -\frac{\varphi(\tau)}{|Q(i\omega,\tau)|^2}, \qquad \tau\in I, \tag{2.40}$$

where ψ,φ are continuous and differentiable functions of τ such that $\psi^2+\varphi^2 = |P(i\omega,\tau)|^4$ and $|Q(i\omega,\tau)|^2 = |P(i\omega,\tau)|^2$ for $\tau\in I$. Hence, we have

$$\begin{aligned}
&\theta(\tau) = \arctan(-\psi(\tau)/\varphi(\tau)) \quad \text{if} \quad \sin\theta(\tau)>0, \cos\theta(\tau)>0;\\
&\theta(\tau) = \pi/2 \quad \text{if} \quad \sin\theta(\tau)=1, \cos\theta(\tau)=0;\\
&\theta(\tau) = \pi+\arctan(-\psi(\tau)/\varphi(\tau)) \quad \text{if} \quad \cos\theta(\tau)<0;\\
&\theta(\tau) = 3\pi/2 \quad \text{if} \quad \sin\theta(\tau)=-1, \cos\theta(\tau)=0;\\
&\theta(\tau) = 2\pi+\arctan(-\psi(\tau)/\varphi(\tau)) \quad \text{if} \quad \sin\theta(\tau)<0, \cos\theta(\tau)>0.
\end{aligned} \tag{2.41}$$

It is easy to see that the function $\theta(\tau)$ defined above is continuous on I. Furthermore $\theta'(\tau)$ is well defined for $\theta(\tau)\in(0,2\pi)$ and it is indeed given by (2.34). Observe that if $\theta(\tau)\neq\pi/2, 3\pi/2$, then $\varphi(\tau)\neq 0$, and (2.34) simply follows from (2.41). When $\theta(\tau)=\pi/2, 3\pi/2$, we have $\varphi(\tau)=0$. In this case, we compute $\theta'(\tau)$ directly from (2.40) and obtain

$$-(\sin(\theta(\tau)))\theta'(\tau) = (-\varphi(\tau)/(\psi^2(\tau)+\varphi^2(\tau))^{1/2})'. \tag{2.42}$$

Since $\sin\theta(\tau)\neq 0$ (i.e., $\psi(\tau)\neq 0$), it is easy to see that (2.42) implies (2.34) as well.

Therefore, if $\theta(\tau)\in(0,2\pi), \tau\in I$, then $\theta(\tau)$ is continuous and differentiable. If in addition, $\omega(\tau)$ is positive, continuous, and differentiable on I, then functions $\tau_n(\tau)$ and $S_n(\tau), n\in\mathbf{N}_0$, are all continuous and differentiable.

Remark 2.2. Instead of looking for zeros of S_n, we can look for the zeros of, say, $Z_n = \omega S_n = \omega\tau - \theta(\tau) - 2n\pi = Z_0 - 2n\pi$. Since $\omega>0$, they have the same zeros, and all the functions Z_n have the same shape as Z_0 (they are simply shifted down by $2n\pi$). Furthermore it is easy to check that $\operatorname{sign}(S'_n) = \operatorname{sign}(Z'_n)$ when considering the derivative with respect to τ at the same zero as S_n and Z_n. In most cases of applications, we can assume that $I=[0,\tau_1)$ with $\omega(0)>0$ and $\omega(\tau)\to 0$ as $\tau\to\tau_1$. Then clearly $Z_n(0)<0$ and $Z_n(\tau)<0$ as $\tau\to\tau_1$. Hence either Z_n is negative or it has an even number of zeros (taking into account multiplicity of the zeros).

3. First order characteristic equation.

In this section, we consider the first order characteristic equation

$$D(\lambda,\tau)=0, \tag{3.1}$$

where

$$D(\lambda,\tau) = a(\tau)\lambda + b(\tau) + c(\tau)e^{-\lambda\tau}, \tag{3.2}$$

which belongs to the general class

$$D(\lambda,\tau) = P_n(\lambda,\tau) + Q_m(\lambda,\tau)e^{-\lambda\tau}, \tag{3.3}$$

where P_n, Q_m are polynomials in λ with $n > m$. In our case $P(\lambda,\tau) := P_n(\lambda,\tau) = a(\tau)\lambda + b(\tau)$ is a first order polynomial in λ, $Q(\lambda,\tau) := Q_m(\lambda,\tau) = c(\tau)$. The coefficients a, b, c are real smooth functions of τ assumed to have continuous derivatives in τ and

$$b(\tau) + c(\tau) \neq 0 \quad \forall\tau \geq 0. \tag{3.4}$$

Due to assumption (3.4), $\lambda = 0$ cannot be a root of (3.1) and a stability switch (or a cross of the imaginary axis) necessarily occurs with $\lambda = \pm i\omega$ with $\omega > 0$. Without loss of generality assume $\lambda = i\omega$, $\omega > 0$, as a root of (3.2). Hence at $\lambda = i\omega$ we have

$$P(i\omega,\tau) = b(\tau) + i\omega a(\tau), \quad Q(i\omega,\tau) = c(\tau), \tag{3.5}$$

i.e.,

$$F(\omega,\tau) = |P(i\omega,\tau)|^2 - |Q(i\omega,\tau)|^2 = \omega^2a^2 + b^2 - c^2, \tag{3.6}$$

from which $F(\omega,\tau) = 0$ gives a solution for $\omega(\tau) > 0$:

$$\omega(\tau) = \left(\frac{c^2(\tau) - b^2(\tau)}{a^2(\tau)}\right)^{1/2}, \tag{3.7}$$

which is defined if $|c(\tau)| > |b(\tau)|$ and $a(\tau) \neq 0$. Furthermore, since $P_R(i\omega,\tau) = b(\tau)$, $P_I(i\omega,\tau) = \omega a(\tau)$, $Q_R(i\omega,\tau) = c(\tau)$, $Q_I(i\omega,\tau) = 0$, (2.16) give

$$\sin\theta(\tau) = \frac{\omega(\tau)a(\tau)}{c(\tau)}, \quad \cos\theta(\tau) = -\frac{b(\tau)}{c(\tau)}. \tag{3.8}$$

Let

$$I = \{\tau : \tau \geq 0, a(\tau) \neq 0 \quad \text{and} \quad |b(\tau)| < |c(\tau)|\}.$$

Let $\theta(\tau) \in I$ be the solution of (3.8). Then a stability switch may occur, through the roots $\lambda = \pm i\omega(\tau)$, where $\omega(\tau) > 0$ is given by (3.7), at the τ values

$$\tau_n = \frac{\theta(\tau) + n2\pi}{\omega(\tau)} \tag{3.9}$$

for $n \in \mathbf{N}_0 := \{0, 1, 2, \ldots\}$. Then for each $n \in \mathbf{N}_0$ (3.9) defines the maps $\tau_n : I \rightarrow \mathbf{R}_{+0}$, and the stability switch may occur only for the τ values at which

$$\tau_n(\tau) = \tau \quad \text{for some} \quad n \in \mathbf{N}_0. \tag{3.10}$$

Hence (3.7), (3.8) define the maps (3.9), and the occurrence of stability switches takes place at the zeros of the functions

$$S_n(\tau) := \tau - \tau_n(\tau), \quad n \in \mathbf{N}_0. \tag{3.11}$$

Remark 3.1. We remark here that for $\tau \in I, \theta(\tau)$ is continuous and differentiable in τ. To see this, we observe that due to (3.4), we have $\cos\theta(\tau) \neq 1$ for $\tau \in I$. Hence for $\tau \in I,\ \theta(\tau) \neq 0, 2\pi$. This shows that $\theta(\tau)$ is continuous and differentiable. As a result, we see that $\tau_n(\tau)$ are also continuous and differentiable for $\tau \in I$. Now we want to see if it is possible to determine the direction in which the pair of imaginary roots $\lambda = \pm i\omega(\tau^*)$ (where τ^* is such that $S_n(\tau^*) = 0$ for some n) crosses the imaginary axis as τ increases. In view of the fact that it is now quite straightforward to generate the graphs of $S_n(\tau)$ by popular software such as Maple, we want to connect the rather abstract value of

$$R(\tau) := \text{sign}\left\{\left.\frac{d\text{Re}\lambda}{d\tau}\right|_{\lambda = i\omega(\tau^*)}\right\}$$

with the intuitive and easy to use one

$$S(\tau) := \text{sign}\left\{\left.\frac{dS_n(\tau)}{d\tau}\right|_{\tau=\tau^*}\right\} = \text{sign}\left\{1 - \left.\frac{d\tau_n(\tau)}{d\tau}\right|_{\tau=\tau^*}\right\}. \tag{3.12}$$

Observe that

$$F'_\omega(\omega, \tau) = 2\omega a^2 > 0 \tag{3.13}$$

since $\omega(\tau) > 0$. Therefore, (2.21) in Theorem 2.2 becomes

$$\text{sign}\left\{\left.\frac{d\text{Re }\lambda}{d\tau}\right|_{\lambda=i\omega(\tau^*)}\right\} = \text{sign}\left\{\left.\frac{dS_n(\tau)}{d\tau}\right|_{\tau=\tau^*}\right\}. \tag{3.14}$$

Hence, we have the following.

Theorem 3.1. *The characteristic equation* (3.1) *admits a pair of simple and conjugate roots* $\lambda_+(\tau^*) = i\omega(\tau^*), \lambda_-(\tau^*) = -i\omega(\tau^*)$, $\omega(\tau^*) > 0$, *at* $\tau^* \in I$ *if* $S_n(\tau^*) = 0$, *for some* $n \in \mathbf{N}_0$. *This pair of simple conjugate pure imaginary roots crosses the imaginary axis from left to right if* $\delta(\tau^*) > 0$ *and crosses the imaginary axis from right to left if* $\delta(\tau^*) < 0$, *where*

$$\delta(\tau^*) = \text{sign}\left\{\left.\frac{d\text{Re}\lambda}{d\tau}\right|_{\lambda=i\omega(\tau^*)}\right\} = \text{sign}\left\{\left.\frac{dS_n(\tau)}{d\tau}\right|_{\tau=\tau^*}\right\}. \tag{3.15}$$

The following analytical result on $R(\tau)$ is useful for determining analytically the τ values at which a stability switch occurs.

Theorem 3.2. *For the characteristic equation* (3.1),

$$\text{sign}\left\{\left.\frac{d\text{Re}\lambda}{d\tau}\right|_{\lambda=i\omega(\tau^*)}\right\} = \text{sign}\,\{a^2(\tau)\omega(\tau)\omega'(\tau)(a(\tau)b(\tau) + c(\tau)^2\tau) + \omega^2(\tau)a^2(\tau)(a'(\tau)b(\tau) - a(\tau)b'(\tau) + c^2(\tau))\}. \tag{3.16}$$

Proof. Denote by a', b', c' the derivatives of $a(\tau), b(\tau), c(\tau)$ with respect to τ. Differentiating (3.1) with respect to τ we obtain

$$\frac{d\lambda}{d\tau} = \frac{\lambda c(\tau)e^{-\lambda\tau} - (a'(\tau)\lambda + b'(\tau) + c'(\tau)e^{-\lambda\tau})}{a(\tau) - c(\tau)\tau e^{-\lambda\tau}}. \tag{3.17}$$

It is convenient to consider $(d\lambda/d\tau)^{-1}$. Hence, from (3.17) we have

$$\left(\frac{d\lambda}{d\tau}\right)^{-1} = \frac{a(\tau)e^{\lambda\tau} - c(\tau)\tau}{\lambda c(\tau) - (a'(\tau)\lambda + b'(\tau))e^{\lambda\tau} - c'(\tau)}, \tag{3.18}$$

where, due to (3.1)

$$e^{\lambda\tau} = -\frac{c(\tau)}{a(\tau)\lambda + b(\tau)}. \tag{3.19}$$

Therefore, substituting (3.19) in (3.18), we have

$$\left(\frac{d\lambda}{d\tau}\right)^{-1} = \frac{-\frac{a(\tau)c(\tau)}{a(\tau)\lambda+b(\tau)} - c(\tau)\tau}{\lambda c(\tau) + \frac{c(\tau)(a'(\tau)\lambda+b'(\tau))}{a(\tau)\lambda+b(\tau)} - c'(\tau)}. \tag{3.20}$$

Now, we compute (3.20) at $\lambda = i\omega(\tau)$. We have

$$\left(\frac{d\lambda}{d\tau}\right)^{-1}\Bigg|_{\lambda=i\omega(\tau)} = \frac{-\frac{ac(b-i\omega a)}{(\omega^2a^2+b^2)} - c\tau}{i\omega c + \frac{c(i\omega a'+b')(b-i\omega a)}{\omega^2a^2+b^2} - c'}. \tag{3.21}$$

Since $\omega^2(\tau)a^2(\tau) + b^2(\tau) = c^2(\tau)$ for any $\tau \in I$, we obtain

$$\left(\frac{d\lambda}{d\tau}\right)^{-1}\Bigg|_{\lambda=i\omega(\tau)} = \frac{-(ab + c^2\tau) + i\omega a^2}{\omega^2aa' + bb' - cc' + i\omega(a'b - ab' + c^2)}. \tag{3.22}$$

Since $F(\omega,\tau) = 0$ for all $\tau \in I$, we obtain

$$-\omega\omega' a^2 = \omega^2 aa' + bb' - cc' \tag{3.23}$$

for all $\tau \in I$, which substituted in (3.22) provides

$$\left(\frac{d\lambda}{d\tau}\right)^{-1}\Bigg|_{\lambda=i\omega(\tau)} = \frac{-(a(\tau)b(\tau) + c(\tau)^2\tau) + i\omega(\tau)a(\tau)^2}{-a^2(\tau)\omega(\tau)\omega'(\tau) + i\omega(\tau)(a'(\tau)b(\tau) - a(\tau)b'(\tau) + c^2(\tau))}. \tag{3.24}$$

Therefore, we have

$$\begin{aligned} R(\tau) = \operatorname{sign}\ \{&a^2(\tau)\omega(\tau)\omega'(\tau)(a(\tau)b(\tau) + c(\tau)^2\tau) + \omega^2(\tau)a^2(\tau)(a'(\tau)b(\tau) \\ &-a(\tau)b'(\tau) + c^2(\tau))\}, \end{aligned} \tag{3.25}$$

proving the theorem. □

Example. As an example of first order characteristic equations with delay dependent coefficients, we consider the first model with time delay (simpler one) introduced by Bence and Nisbet [3] for a population of sessile invertebrates. (This population was previously studied by Roughgarden, Iwasa, and Baxter [25] and Roughgarden and Iwasa [24] in terms of different mathematical models.) This model is a two-stage model in which the population is divided into an adult population, which is explicitly modeled, and a juvenile population which is modeled implicitly. The model takes the form

$$\frac{dA}{dt} = se^{-m_J\tau}\max\{0, 1 - a_AA(t-\tau)\} - m_AA(t), \tag{3.26}$$

where $A(t)$ represents the adult population, s is the settlement rate of juveniles, $e^{-m_J\tau}$ is the through-stage survival probability of juveniles, m_A is the mortality rate of adults, and τ is the fixed time delay between settlement and recruitment into the adult population. Finally $a_A > 0$ is the amount of space occupied by an adult individual. The characteristic equation at steady state takes the form (3.1) with

$$a(\tau) = 1, \quad b(\tau) = m_A, \quad c(\tau) = a_A s e^{-m_J \tau}, \tag{3.27}$$

and $\tau \in \mathbf{R}_{+0}$. Furthermore,

$$b(0) + c(0) = m_A + a_A s > 0,$$

thus ensuring that at $\tau = 0$ we have one negative eigenvalue.

Let $\omega(\tau)$ be the positive solution of

$$\omega^2(\tau) = a_A^2 s^2 e^{-2m_J \tau} - m_A^2 \tag{3.28}$$

which exists provided that

$$a_A s > m_A, \quad \tau < \tau_1 := \frac{1}{m_J} \log\left(\frac{a_A s}{m_A}\right). \tag{3.29}$$

Then, eigenvalues $\lambda_+ = i\omega(\tau), \lambda_- = -i\omega(\tau)$, $\omega(\tau) > 0$, can only occur for delays τ in the interval $I = (0, \tau_1)$. No stability switches for $\tau \geq \tau_1$. It is interesting to determine the values of τ at which $R(\tau) = 1$ and those at which $R(\tau) = -1$. To this end we can use formula (3.16)

$$R(\tau) = \operatorname{sign}\,\{a^2 \omega\omega'(ab + c^2\tau) + \omega^2 a 2(a'b - ab') + \omega^2 c^2\},$$

where a, b, c are coefficients in (3.27) and ω is defined by (3.28). From (3.27) we see that $a' = b' = 0$, $ab = m_A$,

$$\omega\omega' = -m_J a_A^2 s^2 e^{-2m_J\tau} = -m_J c^2. \tag{3.30}$$

Substituting (3.30) in (3.16) we obtain

$$\begin{aligned} R(\tau) &= \operatorname{sign}\,\{-m_J c^2(m_A + c^2\tau) + \omega^2 c^2\} \\ &= \operatorname{sign}\,\{-m_A(m_J + m_A) - c^2(m_J\tau - 1)\}. \end{aligned} \tag{3.31}$$

Then if $(1/m_J) < \tau_1$, (3.31) shows that in the interval $((1/m_J), \tau_1)$ we have $R(\tau) = -1$, i.e., possible stability switches can only occur toward stability. Assume that $\tau < \min\{(1/m_J), \tau_1\}$. Then (3.31) takes the form

$$R(\tau) = \operatorname{sign}\,\{(1 - m_J\tau) a_A^2 s^2 e^{-2m_J\tau} - m_A(m_J + m_A)\}, \tag{3.32}$$

which yields the following conclusions:

(i) If the parameters satisfy

$$m_A < a_A s < m_A \left(1 + \frac{m_J}{m_A}\right)^{1/2}, \tag{3.33}$$

then for all $\tau \in [0, \tau_1)$ we have $R(\tau) = -1$. In such cases, a stability switch from unstable to stable may occur. Since at $\tau = 0$ the steady state is asymptotically stable, then it remains asymptotically stable for all $\tau \in [0, \tau_1)$.

(ii) Assume that the parameters satisfy

$$a_A s > m_A \left(1 + \frac{m_J}{m_A}\right)^{1/2}. \tag{3.34}$$

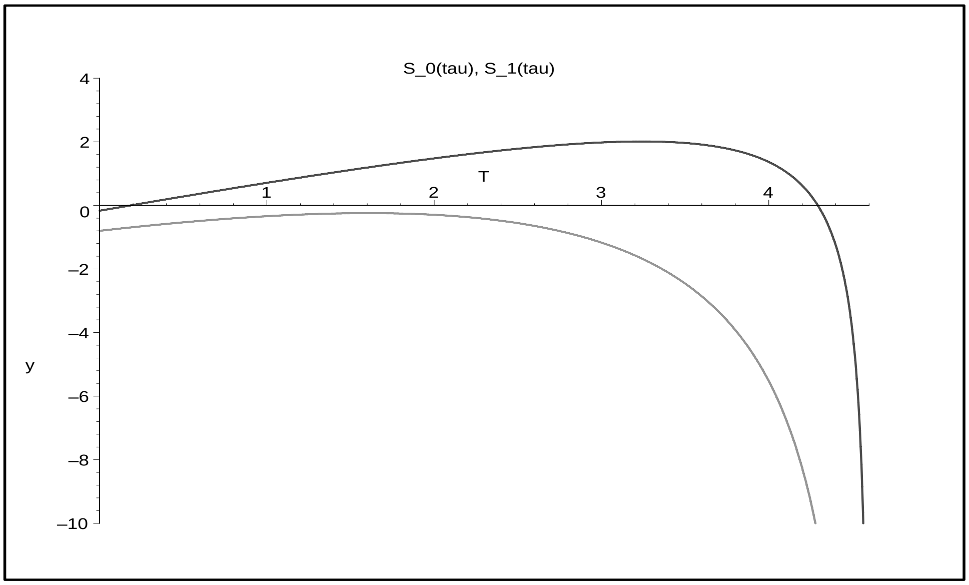


FIG. 1. *Graph of stability switch in terms of time delay for the first model of Bence and Nisbet* [3]. *The top curve is* $S_0(\tau)$.

Then there exists a τ_c, $0 < \tau_c < \tau_1$, such that

$$\operatorname{sign}\left\{ \frac{d\mathrm{Re}\,\lambda}{d\tau}\bigg|_{\lambda=i\omega(\tau)} \right\} > 0$$

in $(0, \tau_c)$, vanishes at τ_c, and is negative in (τ_c, τ_1). τ_c is the unique zero of

$$\varphi(\tau) := (1 - m_J\tau)a_A^2 s^2 e^{-2m_J\tau} - m_A(m_J + m_A), \quad \tau \in (0, \tau_1). \tag{3.35}$$

The statements of (i), (ii) are helpful in choosing the parameters to perform relevant numerical simulations. In Figure 1, we plot the graph of the map $S_0(\tau)$ versus τ in the interval $I = [0, \tau_1)$ for the following set of parameters satisfying (3.34):

$$a_A = 1, \quad m_A = 1, \quad m_J = 0.5, \quad s = 10,$$

with $\tau_1 = 4.605$. The graph of $S_0(\tau)$ versus τ in Figure 1 shows that $S_0(\tau)$ has two zeros, the first at the value $\tau_{01} = 0.20$, the second at the value $\tau_{02} = 4.24$, and $S_1(\tau) < 0$ on $(0, \tau_1)$. According to Theorem 3.1 at τ_{01} a stability switch occurs toward instability whereas at τ_{02} the stability switch occurs toward stability. Hence, for the model by Bence and Nisbet [3], as confirmed by other computer simulations, intermediate delays ($\tau \in (0.2, 4.24)$) show a destabilizing effect on the steady state, whereas large delays ($\tau > 4.24$) have a stabilizing one. For $\tau \in (\tau_{01}, \tau_{02})$, the steady state is unstable whereas it is asymptotically stable for $0 \le \tau < \tau_{01}$ and for any $\tau > \tau_{02}$. These results are in agreement with our computer simulations using XPP.

4. Second order characteristic equation. In this section, we consider the characteristic equation

$$D(\lambda, \tau) := \lambda^2 + a(\tau)\lambda + b(\tau)\lambda e^{-\lambda\tau} + c(\tau) + d(\tau)e^{-\lambda\tau} = 0; \tag{4.1}$$

$\tau \in \mathbf{R}_{+0}$ and $a(\tau), b(\tau), c(\tau), d(\tau) : \mathbf{R}_{+0} \to \mathbf{R}$ are differentiable functions of class $C^1(\mathbf{R}_{+0})$ such that $c(\tau) + d(\tau) \neq 0$ for all $\tau \in \mathbf{R}_{+0}$, and for any τ, $b(\tau), d(\tau)$ are not simultaneously zero. We have

$$P(\lambda, \tau) := P_n(\lambda, \tau) = \lambda^2 + a(\tau)\lambda + c(\tau), \qquad Q(\lambda, \tau) := Q_m(\lambda, \tau) = b(\tau)\lambda + d(\tau).$$

We assume that $P_n(\lambda, \tau)$ and $Q_m(\lambda, \tau)$ cannot have common imaginary roots. That is, for any real number ω,

$$P_n(i\omega, \tau) + Q_m(i\omega, \tau) \neq 0. \tag{4.2}$$

We have

$$F(\omega, \tau) = |P(i\omega, \tau)|^2 - |Q(i\omega, \tau)|^2 = (c - \omega^2)^2 + \omega^2 a^2 - (\omega^2 b^2 + d^2). \tag{4.3}$$

Hence, $F(\omega, \tau) = 0$ implies

$$\omega^4 - \omega^2(b^2 + 2c - a^2) + (c^2 - d^2) = 0, \tag{4.4}$$

and its roots are given by

$$\omega_+^2 = \frac{1}{2}\{(b^2 + 2c - a^2) + \Delta^{1/2}\}, \quad \omega_-^2 = \frac{1}{2}\{(b^2 + 2c - a^2) - \Delta^{1/2}\}, \tag{4.5}$$

where

$$\Delta = (b^2 + 2c - a^2)^2 - 4(c^2 - d^2). \tag{4.6}$$

Therefore, the following holds:

$$2\omega_\pm^2 - (b^2 + 2c - a^2) = \pm\Delta^{1/2}. \tag{4.7}$$

Furthermore, $P_R(i\omega, \tau) = c(\tau) - \omega^2(\tau)$, $P_I(i\omega, \tau) = \omega(\tau)a(\tau)$, $Q_R(i\omega, \tau) = d(\tau)$, $Q_I(i\omega, \tau) = \omega(\tau)b(\tau)$. Hence (2.16) becomes

$$\sin\theta(\tau) = \frac{-(c - \omega^2)\omega b + \omega a d}{\omega^2 b^2 + d^2}, \quad \cos\theta(\tau) = -\frac{(c - \omega^2)d + \omega^2 ab}{\omega^2 b^2 + d^2}, \tag{4.8}$$

which jointly with (4.4) defines the maps (2.19). Now, from (4.3) we have

$$\begin{aligned} F'_\omega(\omega, \tau) &= 2(c - \omega^2)(-2\omega) + 2\omega a^2 - 2\omega b^2 \\ &= 2\omega[2\omega^2 - (b^2 + 2c - a^2)] \\ &= 2\omega_\pm[\pm\Delta^{1/2}], \end{aligned} \tag{4.9}$$

where $\omega_\pm(\tau) > 0$. Hence (2.21) in Theorem 2.2 becomes

$$\operatorname{sign}\left\{\left.\frac{d\mathrm{Re}\,\lambda}{d\tau}\right|_{\lambda = i\omega_\pm}\right\} = \operatorname{sign}\{\pm\Delta^{1/2}\}\operatorname{sign}\left\{\left.\frac{dS_n(\tau)}{d\tau}\right|_{\tau = \tau^*}\right\}. \tag{4.10}$$

This proves the following theorem.

THEOREM 4.1. *The characteristic equation* (4.1) *has a pair of simple and conjugate pure imaginary roots $\lambda = \pm i\omega(\tau^*)$, $\omega(\tau^*)$ real, at $\tau^* \in I$ if $S_n(\tau^*) = \tau^* - \tau_n(\tau^*) = 0$ for some $n \in \mathbf{N}_0$. If $\omega(\tau^*) = \omega_+(\tau^*)$, this pair of simple conjugate pure imaginary* *roots crosses the imaginary axis from left to right if* $\delta_+(\tau^*) > 0$ *and crosses the imaginary axis from right to left if* $\delta_+(\tau^*) < 0$, *where*

$$\delta_+(\tau^*) := \operatorname{sign}\left\{ \left.\frac{d\mathrm{Re}\ \lambda}{d\tau}\right|_{\lambda = i\omega_+(\tau^*)} \right\} = \operatorname{sign}\left\{ \left.\frac{dS_n(\tau)}{d\tau}\right|_{\tau=\tau^*} \right\}. \tag{4.11}$$

If $\omega(\tau^*) = \omega_-(\tau^*)$, *this pair of simple conjugate pure imaginary roots crosses the imaginary axis from left to right if* $\delta_-(\tau^*) > 0$ *and crosses the imaginary axis from right to left if* $\delta_-(\tau^*) < 0$, *where*

$$\delta_-(\tau^*) := \operatorname{sign}\left\{ \left.\frac{d\mathrm{Re}\ \lambda}{d\tau}\right|_{\lambda = i\omega_-(\tau^*)} \right\} = -\operatorname{sign}\left\{ \left.\frac{dS_n(\tau)}{d\tau}\right|_{\tau=\tau^*} \right\}. \tag{4.12}$$

We remark that if $\omega_+(\tau^*) = \omega_-(\tau^*) = \omega(\tau^*)$, then $\Delta(\tau^*) = 0$ and

$$\operatorname{sign}\left\{ \left.\frac{d\mathrm{Re}\ \lambda}{d\tau}\right|_{\lambda = i\omega(\tau^*)} \right\} = 0. \tag{4.13}$$

The same is true when $S_n'(\tau^*) = 0$. The following result can be useful in identifying values of τ where stability switches may take place. (In using this theorem, the plus or minus signs are to be used consistently on both sides of the equations.)

THEOREM 4.2. *Assume that for all* $\tau \in I$, $\omega(\tau)$ *is defined as a solution of* (4.4). *Then*

$$\delta_\pm(\tau) = \operatorname{sign}\{\pm\Delta^{1/2}(\tau)\}\operatorname{sign}D_\pm(\tau), \tag{4.14}$$

where

$$\begin{aligned} D_\pm(\tau) = {} & \omega_\pm^2[(\omega_\pm^2 b^2 + d^2) + a'(c - \omega_\pm^2) + bd' - b'd - ac'] \\ & + \omega_\pm\omega_\pm'[\tau(\omega_\pm^2 b^2 + d^2) - bd + a(c - \omega_\pm^2) + 2\omega_\pm^2 a] \end{aligned}$$

for all $\tau \in I$.

Proof. Let us differentiate with respect to τ the characteristic equation (4.1). We obtain

$$\left(\frac{d\lambda}{d\tau}\right)^{-1} = \frac{-\frac{2\lambda + a}{\lambda^2 + a\lambda + c} + \frac{b}{b\lambda + d} - \tau}{\lambda + \frac{a'\lambda + c'}{\lambda^2 + a\lambda + c} - \frac{b'\lambda + d'}{b\lambda + d}}. \tag{4.15}$$

Now, let $\lambda = i\omega(\tau)$ where $\omega(\tau) > 0$ satisfies (4.4). Hence, we have

$$\begin{aligned} \mathrm{Re}\left(\frac{d\lambda}{d\tau}\right)^{-1}\Bigg|_{\lambda = i\omega} &= \mathrm{Re}\left\{ \frac{-\frac{(2i\omega + a)((c-\omega^2) - i\omega a)}{(c-\omega^2)^2 + \omega^2 a^2} + \frac{b(d - i\omega b)}{d^2 + \omega^2 b^2} - \tau}{i\omega + \frac{(i\omega a' + c')((c-\omega^2) - i\omega a)}{(c-\omega^2)^2 + \omega^2 a^2} - \frac{(i\omega b' + d')(d - i\omega b)}{d^2 + \omega^2 b^2}} \right\} \\ &= \mathrm{Re}\left\{ \frac{[bd - \tau(d^2 + \omega^2 b^2) - 2\omega^2 a - a(c - \omega^2)] + i[\omega(2\omega^2 - (b^2 + 2c - a^2))]}{[\omega^2 aa' + c'(c - \omega^2) - \omega^2 bb' - dd'] + i[\omega((d^2 + \omega^2 b^2) + a'(c - \omega^2) - ac' + bd' - b'd)]} \right\}. \end{aligned} \tag{4.16}$$

Therefore, from (4.16) we have

$$\begin{aligned} \operatorname{sign}\left\{ \mathrm{Re}\left(\frac{d\lambda}{d\tau}\right)^{-1}\Bigg|_{\lambda = i\omega} \right\} = \operatorname{sign}\{ & [\tau(d^2 + \omega^2 b^2) - bd + a(c - \omega^2) + 2\omega^2 a] \\ & \times[-\omega^2 aa' - c'(c - \omega^2) + \omega^2 bb' + dd'] \\ & + \omega^2[2\omega^2 - (b^2 + 2c - a^2)] \\ & \times[(\omega^2 b^2 + d^2) + a'(c - \omega^2) + bd' - b'd - ac']\}. \end{aligned} \tag{4.17}$$

Differentiating both sides of (4.4) with respect to τ we obtain

$$\omega\omega'[2\omega^2 - (b^2 + 2c - a^2)] = -\omega^2 aa' - c'(c - \omega^2) + \omega^2 bb' + dd'. \tag{4.18}$$

Furthermore, $\omega(\tau)$ must be one of the two roots $\omega_\pm(\tau)$ given by (4.5); therefore,

$$2\omega_\pm^2 - (b^2 + 2c - a^2) = \pm\Delta^{1/2}. \tag{4.19}$$

Combining (4.18) and (4.19) in (4.17), we obtain

$$\begin{aligned} &\operatorname{sign}\left\{\operatorname{Re}\left(\frac{d\lambda}{d\tau}\right)^{-1}\bigg|_{\lambda=i\omega_\pm(\tau)}\right\} \\ &= \operatorname{sign}\,\{\pm\Delta^{1/2}(\tau)\} \\ &\quad \times\operatorname{sign}\,\{\omega_\pm^2[(\omega_\pm^2 b^2 + d^2) + a'(c - \omega_\pm^2) + bd' - b'd - ac'] \\ &\qquad + \omega_\pm\omega_\pm'[\tau(\omega_\pm^2 b^2 + d^2) - bd + a(c - \omega_\pm^2) + 2\omega_\pm^2 a]\}. \end{aligned}$$

Since

$$\operatorname{sign}\left\{\frac{d\mathrm{Re}\ \lambda}{d\tau}\right\} = \operatorname{sign}\left\{\mathrm{Re}\ \left(\frac{d\lambda}{d\tau}\right)^{-1}\right\},$$

the proof is completed. □

Remark 4.1. In almost all the application problems that we have encountered so far that have characteristic equations of the form (4.1), often only ω_+ is feasible. In these cases, stability switches occur only at the roots of $S_0^+(\tau) = \tau - \tau_0^+(\tau) = 0$, where

$$\tau_0^+(\tau) = \theta_+(\tau)/\omega_+(\tau), \qquad \tau \in I, \tag{4.20}$$

and $\theta_+(\tau)$ is the solution of (4.8) when $\omega = \omega_+$.

However, if both ω_+ and ω_- are feasible for $\tau \in I$, then we have the following two sequences of functions on I:

$$S_n^+(\tau) = \tau - (\theta_+(\tau) + 2n\pi)/\omega_+(\tau), \quad S_n^-(\tau) = \tau - (\theta_-(\tau) + 2n\pi)/\omega_-(\tau), \tag{4.21}$$

where the notations are self-evident. Clearly $S_n^+(\tau) > S_{n+1}^+(\tau)$ and $S_n^-(\tau) > S_{n+1}^-(\tau)$ for all $n \in \mathbf{N}_0, \tau \in I$. In addition to this, we have the following simple statement.

THEOREM 4.3. *Assume that $S_0^+(\tau) > S_0^-(\tau)$ on I. Then $S_n^+(\tau) > S_n^-(\tau)$ on I for all $n \in \mathbf{N}_0$.*

Proof. It is easy to see that $S_0^+(\tau) > S_0^-(\tau)$ on I implies that

$$\theta_+(\tau)/\omega_+(\tau) < \theta_-(\tau)/\omega_-(\tau), \qquad \tau \in I. \tag{4.22}$$

Note that $\omega_+(\tau) > \omega_-(\tau)$ on I; the theorem follows from (4.21). □

Remark 4.2. When both ω_+ and ω_- are feasible for $\tau \in I$, it is easy to imagine that the stability switches may depend on all real roots of $S_n^+(\tau) = 0$ and $S_n^-(\tau) = 0$. In such situations, one must examine all these possible real roots in order to determine the stability of the equilibrium. To illustrate this, let us consider such a scenario. Assume that Theorem 4.3 holds true and the equilibrium is asymptotically stable when $\tau = 0$. Assume further that $S_0^+(\tau) = 0$ at $t_1 = \tau_{01}^+$ and $t_4 = \tau_{02}^+$, $S_0^-(\tau) = 0$ at $t_2 = \tau_{01}^-$ and $t_3 = \tau_{02}^-$, and no real roots for $S_n^+(\tau) = 0$ and $S_n^-(\tau) = 0$ when $n > 0$. Then it is easy to see (Figure 2) that $t_1 < t_2 < t_3 < t_4$. Careful but simple examination shows

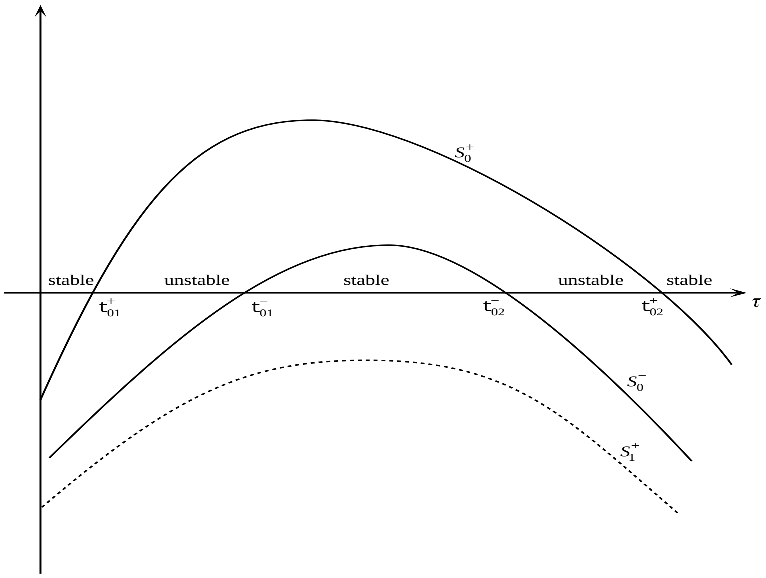


FIG. 2. *Illustration for Remark* 4.2.

that the equilibrium is asymptotically stable for $\tau \in [0, t_1) \cup (t_2, t_3) \cup ((t_4, \infty) \cap I)$ and is unstable for $\tau \in (t_1, t_2) \cup (t_3, t_4)$. More complicated scenarios are clearly conceivable.

Example. As an example for the second order characteristic equations with delay dependent coefficients, we consider the second model with time delay and stage structure introduced by Bence and Nisbet [3] for a population of sessile invertebrates. Again, the model is a two-stage model in which population is divided into adult and juvenile populations, both of which are explicitly modeled. The model takes the form

$$J'(t) = s[F(t) - e^{-m_J\tau} F(t-\tau)] - m_J J(t), \tag{4.23}$$

$$A'(t) = se^{-m_J\tau} F(t-\tau) - m_A A(t), \tag{4.24}$$

$$F(t) = \max\{0, 1 - a_J J(t) - a_A A(t)\}, \tag{4.25}$$

where $sF(t)$ represents the newly settled juveniles and $se^{-m_J\tau}F(t-\tau)$ the ones that become adults. When $a_J = 0$, this model reduces to (3.26). Here, s, m_g, m_A, a_A are positive constants and a_J is a nonnegative constant. This model was systematically studied by Kuang and So in [22]. It should be mentioned here that the last statement of their Theorem 4.1, which says that $S_0^+(\tau) > 0$ (with respect to the positive steady state) implies that the positive steady state is unstable, is not fully justified. To be accurate, it requires the assumption that ω_- is not feasible. The analysis and simulation (with both Maple and XPP) we conducted so far, however, indeed suggest that this assumption may in fact hold for all biologically meaningful parameters.

Let J^*, A^*, F^* denote equilibrium population sizes of juveniles, adults, and equilibrium free space, respectively. We have

$$J^* = \frac{sF^*}{m_J}(1 - e^{-m_J\tau}), \qquad A^* = \frac{sF^*}{m_A}e^{-m_J\tau}. \tag{4.26}$$

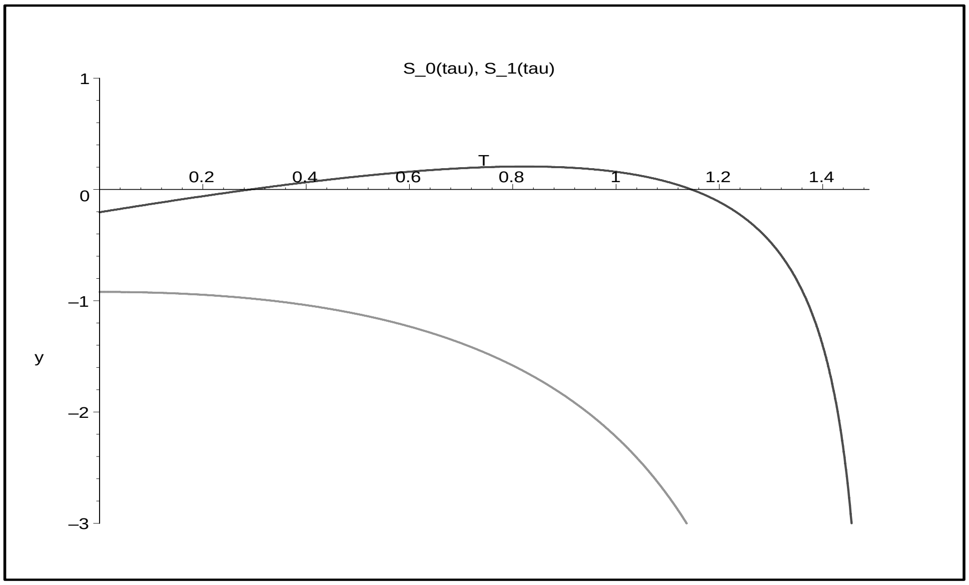


FIG. 3. *Graph of stability switch in terms of time delay for the second model of Bence and Nisbet* [3]*. The top curve is* $S_0(\tau)$*.*

It follows that $F^* \neq 0$ and

$$F^* = 1 - sF^* \left[\frac{a_J}{m_J}(1 - e^{-m_J\tau}) + \frac{a_A}{m_A} e^{-m_J\tau} \right]. \tag{4.27}$$

Following the notation of Bence and Nisbet [3], $\sigma_J = sa_J/m_J, \sigma_A = sa_A/m_A$, we have

$$F^* = [1 + \sigma_J + (\sigma_A - \sigma_J)e^{-m_J\tau}]^{-1}.$$

Thus, system (4.23)–(4.25) has a unique positive steady state (J^*, A^*) and $F^* < 1$. Its characteristic equation at the positive steady state takes the form

$$\lambda^2 + a\lambda + b\lambda e^{-\lambda\tau} + c + de^{-\lambda\tau} = 0, \tag{4.28}$$

where

$$a = m_J + m_A + a_J s, \qquad b = b(\tau) = (a_A - a_J)se^{-m_J\tau}, \tag{4.29}$$

$$c = m_A m_J + a_J m_A s, \qquad d = d(\tau) = (a_A m_J - a_J m_A)se^{-m_J\tau}. \tag{4.30}$$

Clearly, $c + d > 0$, which implies $\lambda = 0$ can never be a root of (4.28). When $\tau = 0$, (4.28) reduces to

$$\lambda^2 + (a + b)\lambda + c + d = 0,$$

which has roots with negative real parts, implying that (J^*, A^*) is locally asymptotically stable when $\tau = 0$.

In Figure 3, we plot the graph of the map $S_0^+(\tau)$ versus τ in the interval $I = [0, \tau_1)$ for the set of parameters satisfying

$$a_A = 1, \quad m_A = 1, \quad a_J = 0.1, \quad m_J = 1, \quad s = 10,$$

with $\tau_1 = 1.49$. It can be shown that for this set of parameters, ω_- is not feasible for $\tau \in I$. The graph of $S_0^+(\tau)$ versus τ in Figure 3 shows that $S_0^+(\tau)$ has two zeros, the first at the value $\tau_{01} = 0.29$, the second at the value $\tau_{02} = 1.14$, whereas $S_1^+(\tau) < 0$ on I. According to Theorem 4.1 and Remark 4.1, we see that at τ_{01} a stability switch occurs toward instability whereas at τ_{02} the stability switch occurs toward stability. Again, we see that intermediate delays show a destabilizing effect on the steady state, whereas large delays have a stabilizing one. For $\tau \in (\tau_{01}, \tau_{02})$, the steady state is unstable, whereas it is asymptotically stable for $0 \le \tau < \tau_{01}$ and for any $\tau > \tau_{02}$. These results are in agreement with our computer simulations using XPP.

5. Discussion.

It is well known that for nonlinear delay systems (including many neutral delay systems) the occurrence of characteristic roots crossing the imaginary axis from left to right as the result of changing certain parameters often ensures the emergence of nontrivial periodic solutions near the steady state as it becomes unstable (Hale and Verduyn Lunel [17]). For well-constructed population models, this scenario is to be expected. XPP simulation can easily confirm this.

We would like to stress here that the geometric criterion presented in the previous section may also be applicable to models with several discrete delays or distributed delays.

Consider, for an example, the following Lotka–Volterra predator prey model with two discrete delays:

$$\begin{cases} x'(t) = x(t)[e_1 - a_1 x(t) - a_2 y(t-\sigma)], \\ y'(t) = y(t)[-e_2 + a_3 x(t-\tau) - a_4 y(t)], \end{cases} \tag{5.1}$$

where all parameters are positive constants. Assume further that it has a positive steady state $E^* = (x^*, y^*)$. Let $r = \sigma/\tau$. The system can be reduced to the following one with dimensionless time t/τ, which, for simplicity, we again denote by t:

$$\begin{cases} x'(t) = \tau x(t)[e_1 - a_1 x(t) - a_2 y(t-r)], \\ y'(t) = \tau y(t)[-e_2 + a_3 x(t-1) - a_4 y(t)]. \end{cases} \tag{5.2}$$

At E^*, the characteristic equation is

$$\lambda^2 + \lambda a^* \tau + b^* \tau^2 + c^* \tau^2 e^{-\lambda(r+1)} = 0, \tag{5.3}$$

where $a^* = a_1 x^* + a_4 y^*,\ b^* = a_1 a_4 x^* y^*,\ c^* = a_2 a_3 x^* y^*$. Since $\tau = \sigma/r$, we see that the characteristic equation takes the form

$$\lambda^2 + \lambda A(r,\sigma) + B(r,\sigma) + C(r,\sigma) e^{-\lambda(r+1)} = 0, \tag{5.4}$$

where $A(r,\sigma) = a^*\sigma/r,\ B(r,\sigma) = b^*\sigma^2/r^2,\ C(r,\sigma) = c^*\sigma^2/r^2$, and r may vary in $\mathbf{R}_+$. Clearly, for each fixed value of delay σ, we have a characteristic equation with parameters dependent on r. Notice that all coefficients of (5.4) are positive, and hence all its real roots must be negative. Stability switching may occur when imaginary roots $\lambda = i\omega$ exist and cross the imaginary axis.

It is straightforward to find that ω must be the solution of

$$F(\omega, r) := \omega^4 + \omega^2(A^2(r) - 2B(r)) + B^2(r) - C^2(r). \tag{5.5}$$

(For simplicity, we drop the dependence of σ.) This gives

$$\omega_\pm^2 = \frac{1}{2}\{2B(r) - A^2(r) \pm [(2B(r) - A^2(r))^2 + 4(C^2(r) - B^2(r))]^{1/2}\}. \tag{5.6}$$

Let $\theta(r) \in [0, 2\pi]$ be the solution of

$$\cos\theta(r) = \frac{\omega^2(r) - B(r)}{C(r)}, \qquad \sin\theta(r) = \frac{\omega(r)A(r)}{C(r)}. \tag{5.7}$$

The imaginary roots $\lambda = \pm i\omega, \omega(r) > 0$ will appear at the r values which are zeros r^* of

$$S_n(r) = r + 1 - \frac{\theta(r) + 2n\pi}{\omega(r)}, \qquad n \in \mathbf{N}_0. \tag{5.8}$$

Hence, once we know such r^*, we know $\tau^* = \sigma/r^*$. This will give us a pair of delay values (τ^*, σ) at which the stability switch may be possible when increasing the value of $r = \sigma/\tau$ while keeping σ fixed. Of course, such analysis can be performed for each σ for which the solutions (5.6) are feasible. A similar procedure can be applied to σ while keeping τ fixed.

As an example for the applicability of our geometric criterion to models with distributed delays, we consider the following model of single species growth:

$$x'(t) = f\left(\int_{-\tau}^{0} e^{ds}x(t+s)ds\right) - g(x(t)), \tag{5.9}$$

where τ is the maximum stage delay and d is the through-stage death rate, both positive constants. The first term accounts for the births due to all the age groups and the second term represents the death rate. Typical assumptions on f and g are given in [20]. This model can also be viewed as a direct extension of the models studied in [10]. Assume that it admits a positive steady state of $x(t) = x^*$. Then

$$f\left(\frac{1}{d}(1 - e^{-d\tau})x^*\right) = g(x^*).$$

Let

$$a := f'\left(\frac{1}{d}(1 - e^{-d\tau})x^*\right), \qquad b := g'(x^*).$$

Then the linearized equation of (5.9) takes the form

$$u'(t) = a\int_{-\tau}^{0} e^{ds}u(t+s)ds - bu(t). \tag{5.10}$$

By differentiating the above equation one more time and making some simple substitution, we have

$$u''(t) + (b+d)u'(t) + (db - a)u(t) + ae^{-d\tau}u(t-\tau) = 0. \tag{5.11}$$

Clearly, this is a special case of (4.1) and our criteria are applicable.

There are several factors that may have contributed to the current prevalence of models containing only delay independent parameters. These include the following: (1) the authors fail to recognize the need to have some of the parameters become delay dependent; (2) the authors think that delay independent parameters can provide a good enough description or approximation of the dynamics; (3) there is simply a lack of mathematical results and methods to deal with models involving delay dependent parameters.

To see why it is easy to overlook the need of introducing delay dependent parameters in a model, let us consider the well-studied Nicholson's blowfly model proposed in [16]:

$$N' = R - D = R(N(t-\tau)) - \delta N = pN(t-\tau)e^{-N(t-\tau)} - \delta N, \tag{5.12}$$

where N is the sexually mature adult blowfly population density, δ is its individual death rate, and R is the recruitment rate. In arriving at the above form of R, Gurney, Blythe, and Nisbet [16] argued that (1) the rate at which eggs are produced depends only on the current size of adult population, (2) each egg takes about τ units of time to become a sexually mature adult, and (3) the probability of a given egg maturing into a viable adult depends only on the number of competitors of the same age. However, since an individual larva will die at a constant or average rate of d, the through-stage survival rate for a larva to adulthood is $e^{-d\tau}$. Thus a more plausible model should take the form (Cooke et al. [10])

$$N' = e^{-d\tau}R(N(t-\tau)) - \delta N = pe^{-d\tau}N(t-\tau)\exp\{-N(t-\tau)\} - \delta N. \tag{5.13}$$

Now, this equation has a parameter $e^{-d\tau}$ that depends on the time delay τ.

To see the importance of introducing delay dependent parameters, we again use models (5.12) and (5.13). It is well known [21] that for (5.12), if a positive steady state exists and this steady state is unstable for $\tau = \tau_0$, then it remains unstable for $\tau > \tau_0$. That is, a large time delay plays a destabilizing role. However, for model (5.13), one sees the opposite [10]. What is often observed (which can be easily verified by XPP) is that there are two threshold values $0 < \tau_0 < \tau_1$ such that the positive steady state is unstable only when $\tau \in [\tau_0, \tau_1]$. Dramatic differences in dynamics provided by models with delay dependent parameters and models with delay independent parameters like that above seem to be the rule rather than the exception. It is also worth mentioning here that applying existing criteria designed for models with delay independent parameters can lead to speculative or even erroneous statements.

Acknowledgments. The authors would like to thank the two referees for their constructive comments and suggestions.

REFERENCES

[1] W. G. Aiello and H. I. Freedman, *A time-delay model of single-species growth with stage structure,* Math. Biosci., 101 (1990), pp. 139–153.

[2] W. G. Aiello, H. I. Freedman, and J. Wu, *Analysis of a model representing stage-structured population growth with state-dependent time delay,* SIAM J. Appl. Math., 52 (1992), pp. 855–869.

[3] J. R. Bence and R. M. Nisbet, *Space-limited recruitment in open systems: The importance of time delays,* Ecology, 70 (1989), pp. 1434–1441.

[4] E. Beretta and Y. Kuang, *Modeling and analysis of a marine bacteriophage infection with latency period,* Nonlinear Anal. Real World Appl., 2 (2001), pp. 35–74.

[5] S. P. Blythe, R. M. Nisbet, W. S. C. Gurney, and N. MacDonald, *Stability switches in distributed delay models,* J. Math. Anal. Appl., 109 (1985), pp. 388–396.

[6] F. G. Boese, *Stability with respect to the delay: On a paper of K. L. Cooke and P. van den Driessche,* J. Math. Anal. Appl., 228 (1998), pp. 293–321.

[7] K. L. Cooke and Z. Grossman, *Discrete delay, distributed delay and stability switches,* J. Math. Anal. Appl., 86 (1982), pp. 592–627.

[8] K. L. Cooke and P. van den Driesche, *On zeros of some transcendental equations,* Funkcial. Ekvac., 29 (1986), pp. 77–90.

[9] K. L. Cooke and P. van den Driesche, *Analysis of an SEIRS epidemic model with two delays,* J. Math. Biol., 35 (1996), pp. 240–260.

[10] K. L. Cooke, P. van den Driessche, and X. Zou, *Interaction of maturation delay and nonlinear birth in population and epidemic models,* J. Math. Biol., 39 (1999), pp. 332–352.

[11] J. M. Cushing, *Integrodifferential Equations and Delay Models in Population Dynamics*, Lecture Notes in Biomath. 20, Springer-Verlag, Berlin, New York, 1977.

[12] J. M. Cushing, *An Introduction to Structured Population Dynamics*, CBMS-NSF Regional Conf. Ser. Appl. Math. 71, SIAM, Philadelphia, PA, 1998.

[13] B. Ermentrout, XPPAUT Ver. 5.5, University of Pittsburgh, Pittsburgh, PA; available online at http://www.math.pitt.edu/∼bard/xpp/xpp.html.

[14] H. I. Freedman and Y. Kuang, *Stability switches in linear scalar neutral delay equations,* Funkcial. Ekvac., 34 (1991), pp. 187–209.

[15] H. I. Freedman and V. S. H. Rao, *The tradeoff between mutual interference and time lags in predator-prey systems,* Math. Biosci., 45 (1983), pp. 991–1004.

[16] W. S. C. Gurney, S. P. Blythe, and R. M. Nisbet, *Nicholson's blowflies revisited,* Nature, 287 (1980), pp. 17–21.

[17] J. Hale and S. M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.

[18] V. A. A. Jansen, R. M. Nisbet, and W. S. C. Gurney, *Generation cycles in stage structured populations,* Bull. Math. Biol., 52 (1990), pp. 375–396.

[19] V. L. Kocic and G. Ladas, *Global Behavior of Nonlinear Difference Equations of Higher Order with Applications*, Kluwer, Dordrecht, 1993.

[20] Y. Kuang, *Global attractivity and periodic solutions in delay differential equations related to models of physiology and population biology*, Japan J. Indust. Appl. Math., 9 (1992), pp. 205–238.

[21] Y. Kuang, *Delay Differential Equations with Applications in Population Dynamics*, Academic Press, Boston, 1993.

[22] Y. Kuang and J. W.-H. So, *Analysis of a delayed two-stage population model with space-limited recruitment,* SIAM J. Appl. Math., 55 (1995), pp. 1675–1696.

[23] R. M. Nisbet, W. S. C. Gurney, and J. A. J. Metz, *Stage structure models applied in evolutionary ecology,* Biomathematics, 18 (1989), pp. 428–449.

[24] J. Roughgarden and Y. Iwasa, *Dynamics of metapopulations with space-limited subpopulations,* Theoret. Population Biol., 29 (1986), pp. 235–261.

[25] J. Roughgarden, Y. Iwasa, and C. Baxter, *Demographic theory for an open marine population with space-limited recruitment,* Ecology, 66 (1985), pp. 54–67.

SIAM J. Math. Anal.
Vol. 33, No. 5, pp. 1166–1198

BLOCH APPROXIMATION IN HOMOGENIZATION AND APPLICATIONS[*]

CARLOS CONCA[†], RAFAEL ORIVE[‡], AND MUTHUSAMY VANNINATHAN[§]

Abstract. The classical problem of homogenization of elliptic operators with periodically oscillating coefficients is revisited in this paper. As is well known, the homogenization process in a classical framework is concerned with the study of asymptotic behavior of solutions u^ε of boundary value problems associated with such operators when the period $\varepsilon > 0$ of the coefficients is small. In a previous work by C. Conca and M. Vanninathan [*SIAM J. Appl. Math.*, 57 (1997), pp. 1639–1659], a new proof of weak convergence as $\varepsilon \to 0$ towards the homogenized solution was furnished using Bloch wave decomposition.

Following the same approach here, we go further and introduce what we call *Bloch approximation*, which will provide energy norm approximation for the solution u^ε. We develop several of its main features. As a simple application of this new object, we show that it contains both the first and second order correctors. Necessarily, the Bloch approximation will have to capture the oscillations of the solution in a sharper way. The present approach sheds new light and offers an alternative for viewing classical results.

Key words. homogenization, Bloch waves, correctors

AMS subject classifications. 35B27, 35A25, 42C30

PII. S0036141001382200

1. Introduction.

In this paper, the classical problem of homogenization of elliptic operators with periodically oscillating coefficients is revisited. As is well known, the homogenization process is concerned with the study of the behavior of solutions u^ε of boundary value problems associated with such operators when the coefficients are periodic with small period $\varepsilon > 0$. For an excellent introduction to this subject, the reader is referred to the book of A. Bensoussan, J.-L. Lions, and G. Papanicolaou [5]. In a previous work by C. Conca and M. Vanninathan [11], a new proof of weak convergence of u^ε towards the homogenized solution u^* was furnished using Bloch wave decomposition. Following the same approach, we go further and introduce what we call *Bloch approximation* of the solution u^ε. As a simple application of this new object, we treat the problem of correctors in homogenization. At this point, it is worthwhile to remark that the homogenized solution u^* is merely the weak limit of solutions u^ε as $\varepsilon \to 0$. The idea behind introducing correctors is to look for terms (called *first order correctors*) which, when added to the homogenized solution, provide an approximation in the energy norm for all ε sufficiently small. *Second order correctors* yield an error estimate in the energy norm of order $O(\varepsilon)$. The main feature of Bloch approximation is that it contains both the first and second order corrector terms. Another important feature is that it is easily computable in principle.

[*] Received by the editors January 17, 2001; accepted for publication (in revised form) September 18, 2001; published electronically April 8, 2002. This work has been partially supported by FONDAP through its Programme on Mathematical-Mechanics.

http://www.siam.org/journals/sima/33-5/38220.html

[†] Departamento de Ingeniería Matemática, Facultad de Ciencias Físicas y Matemáticas, and Centro de Modelamiento Matemático, UMR 2071 CNRS-UChile, Universidad de Chile, Casilla 170/3 - Correo 3, Santiago, Chile (cconca@dim.uchile.cl).

[‡] Departamento de Matemática Aplicada, Facultad de Ciencias Matemáticas, Universidad Complutense de Madrid, 28040 Madrid, España (orive@sunma4.mat.ucm.es).

[§] IISc-TIFR Mathematics Programme, TIFR Center, P.O. Box 1234, Bangalore, 560012, India (vanni@math.tifrbng.res.in).

Historically, a classical way of obtaining such correctors is to work in the physical space and use multiscale expansion of the solution, which was first introduced in the basic book just cited. As we will see, the method of Bloch waves sheds new light and offers an alternative for viewing the classical results. This method naturally leads us to work in the Fourier space, and thus in a framework dual to the one used in L. Tartar [21], whose method is very general and justifies, in particular, the first term obtained via the multiscale expansion. (There are also other methods of justification based on the analysis in physical space; cf. G. Nguetseng [17] and G. Allaire [1].) However, it is important to mention that the Bloch wave method does not presuppose any multiscale structure of the solution; on the contrary, such a structure of the solution will be a consequence of the present method. Although correctors are generally not unique, our approach yields a posteriori the same ones as those obtained in [5].

Bloch waves and their applications are not by any means new. It is a classical tool which has been in use in solid state physics since the pioneering paper of F. Bloch [6]. However, the basic idea was introduced in the mathematical literature much earlier by G. Floquet [13]. For later developments, let us cite the works of F. Odeh and J. Keller [18] and M. Reed and B. Simon [19]. A deep analysis concerning the partial regularity of the spectrum of Schrödinger's equation with periodic potential was carried out by C. Wilcox [22]. Our point of view regards periodic medium as a perturbation of homogeneous ones. In this context, the book by T. Kato [15] provides excellent analysis when the parameter of perturbation is a scalar. We end this rather incomplete list by citing G. Allaire and C. Conca [2], [3], and P. Gérard et al. [14]. We feel it is also appropriate to cite a recent work by M. Avellaneda, L. Berlyand, and J.-F. Clouet [4], in which the Bloch–Floquet approach is used to provide new homogenization results and handles the boundary layer terms for frequency dependent problems. To conclude, let us refer the reader to C. Conca [8] for a more complete general survey on Bloch waves.

Before proceeding further, we mention a word about the notations adopted in what follows. Unless mentioned explicitly, the usual summation convention with respect to the repeated indices is understood. The constants appearing in various estimates independent of ε are generically denoted by c, c_1, c_2, etc. Apart from the usual norms in Sobolev spaces H^1, H^2, we will also use the following seminorms:

$$|v|_{H^1} = \left\{ \sum_{j=1}^{N} \left\| D_j v \right\|_{L^2}^2 \right\}^{\frac{1}{2}}, \quad |v|_{H^2} = \left\{ \sum_{j,k=1}^{N} \left\| D_{j,k}^2 v \right\|_{L^2}^2 \right\}^{\frac{1}{2}}.$$

Now let us introduce the problem to be studied in this work. We consider the operator

$$A \overset{\text{def}}{=} -\frac{\partial}{\partial y_k}\left(a_{k\ell}(y) \frac{\partial}{\partial y_\ell} \right), \quad y \in \mathbb{R}^N, \tag{1.1}$$

where the coefficients satisfy

$$\begin{cases} a_{k\ell} \in L^\infty_\#(Y), \quad \text{where} \quad Y =]0, 2\pi[^N, \text{ i.e., each } a_{k\ell} \text{ is a} \\ Y\text{-periodic bounded measurable function defined on } \mathbb{R}^N, \text{ and} \\ \exists \alpha > 0 \quad \text{such that} \quad a_{k\ell}(y)\eta_k\eta_\ell \geq \alpha|\eta|^2 \quad \forall \eta \in \mathbb{R}^N, \ y \in Y \text{ a.e.}, \\ a_{k\ell} = a_{\ell k} \quad \forall k, \ell = 1, \ldots, N. \end{cases} \tag{1.2}$$

For each $\varepsilon > 0$, we consider also the operator A^ε, where

$$A^\varepsilon \overset{\text{def}}{=} -\frac{\partial}{\partial x_k}\left(a^\varepsilon_{k\ell}(x)\frac{\partial}{\partial x_\ell}\right) \quad \text{with} \quad a^\varepsilon_{k\ell}(x) = a_{k\ell}\left(\frac{x}{\varepsilon}\right), \quad x \in \mathbb{R}^N. \tag{1.3}$$

In homogenization theory, it is usual to refer to x and y, the slow and the fast variables, respectively. They are related by $y = \frac{x}{\varepsilon}$. Associated with A^ε, let us consider the boundary value problem

$$A^\varepsilon u^\varepsilon = f \quad \text{in} \quad \Omega, \quad u^\varepsilon \in H^1_0(\Omega), \tag{1.4}$$

which is posed in an arbitrary bounded domain Ω in $\mathbb{R}^N$ and where f is a given element in $L^2(\Omega)$. It is classical that the above problem admits one and only one solution.

From the classical work [5], it is known that one can associate to A^ε a homogenized operator A^* given by

$$A^* \overset{\text{def}}{=} -\frac{\partial}{\partial x_k}\left(q_{k\ell}\frac{\partial}{\partial x_\ell}\right). \tag{1.5}$$

The homogenized coefficients $q_{k\ell}$ are constants and their definition is given below. The solution u^ε of (1.4) converges weakly in $H^1_0(\Omega)$ to the so-called homogenized solution u^* characterized by

$$A^* u^* = f \quad \text{in} \quad \Omega, \quad u^* \in H^1_0(\Omega). \tag{1.6}$$

In the present paper, we do not consider the effects of boundaries, postponing them to a subsequent article [9]. In the case of $\mathbb{R}^N$, it is natural to replace the operator A^ε by $(A^\varepsilon + I)$. In that case, if w^ε satisfies

$$\left\{\begin{array}{rll} (A^\varepsilon + I)w^\varepsilon = g & \text{in} & \mathbb{R}^N, \\ w^\varepsilon \rightharpoonup w^* & \text{in} & H^1(\mathbb{R}^N)\text{-weak}, \end{array}\right. \tag{1.7}$$

where g is a given function in $L^2(\mathbb{R}^N)$, then it can be seen that (see Proposition 6.1 below)

$$w^\varepsilon \to w^* \quad \text{in} \quad L^2(\mathbb{R}^N)\text{-strong}. \tag{1.8}$$

In view of the above result, there is no concentration of L^2-energy at infinity, and therefore throughout this paper we will consider a sequence u^ε and a function $f \in L^2(\mathbb{R}^N)$ satisfying

$$\left\{\begin{array}{rll} A^\varepsilon u^\varepsilon = f & \text{in} & \mathbb{R}^N, \\ u^\varepsilon \rightharpoonup u^* & \text{in} & H^1(\mathbb{R}^N)\text{-weak}, \\ u^\varepsilon \to u^* & \text{in} & L^2(\mathbb{R}^N)\text{-strong}. \end{array}\right. \tag{1.9}$$

The central issue in the analysis of the first order correctors is to obtain functions $u^\varepsilon_1 \in H^1(\mathbb{R}^N)$, which can be easily constructed and have the following characteristic property:

$$\|u^\varepsilon - u^* - \varepsilon u^\varepsilon_1\|_{H^1(\mathbb{R}^N)} \to 0 \quad \text{as} \quad \varepsilon \to 0. \tag{1.10}$$

By definition, second order correctors $u_2^\varepsilon \in H^1(\mathbb{R}^N)$ will enjoy the property

$$\|u^\varepsilon - u^* - \varepsilon u_1^\varepsilon - \varepsilon^2 u_2^\varepsilon\|_{H^1(\mathbb{R}^N)} \leq c\varepsilon. \tag{1.11}$$

One of the purposes of this article is to carry out a more general construction than the classical one for correctors, namely, Bloch approximation θ^ε, which contains all the above correctors and justifies the procedure. Apart from this, θ^ε contains a lot of information about the periodic medium which will be amply demonstrated in this paper.

1.1. Survey of the previous results. In the classical book [5] the authors obtain an asymptotic expansion (with $y = \frac{x}{\varepsilon}$) of the form

$$\begin{aligned} u^\varepsilon(x) = \quad & u^*(x) + \varepsilon\left\{\chi_k(y)\frac{\partial u^*}{\partial x_k}(x) + \widetilde{u}_1(x)\right\} \\ & + \varepsilon^2\left\{\chi_{k\ell}(y)\frac{\partial^2 u^*}{\partial x_k x_\ell}(x) + \chi_\ell(y)\frac{\partial \widetilde{u}_1}{\partial x_\ell}(x) + \widetilde{u}_2(x)\right\} + \cdots . \end{aligned} \tag{1.12}$$

Here, χ_k is the unique solution of the cell problem

$$\begin{cases} A\chi_k = \dfrac{\partial a_{k\ell}}{\partial y_\ell} \qquad \text{in} \quad \mathbb{R}^N, \\ \chi_k \in H^1_\#(Y), \quad \mathcal{M}_Y(\chi_k) \overset{\text{def}}{=} \dfrac{1}{|Y|}\displaystyle\int_Y \chi_k dy = 0. \end{cases} \tag{1.13}$$

The function $\chi_{k\ell}$ is characterized as the unique solution of

$$\begin{cases} A\chi_{k\ell} = a_{k\ell} + a_{km}\dfrac{\partial \chi_\ell}{\partial y_m} - \dfrac{\partial}{\partial y_m}(a_{mk}\chi_\ell) - \mathcal{M}_Y(a_{k\ell}) - \mathcal{M}_Y\left(a_{km}\dfrac{\partial\chi_\ell}{\partial y_m}\right) \text{ in } \mathbb{R}^N, \\ \chi_{k\ell} \in H^1_\#(Y), \quad \mathcal{M}_Y(\chi_{k\ell}) = 0. \end{cases} \tag{1.14}$$

Further, $\widetilde{u}_1(x), \widetilde{u}_2(x), \ldots$ are independent of ε and satisfy equations of the type $A^*\widetilde{u}_j = \widetilde{g}_j$ in $\mathbb{R}^N$, where, for instance, $\widetilde{g}_1(x) = b_{jk\ell}D^3_{jk\ell}u^*$, where $b_{jk\ell}$ are constants:

$$b_{jk\ell} = \mathcal{M}_Y\left(a_{jm}\frac{\partial \chi_{k\ell}}{\partial y_m} + a_{k\ell}\chi_j\right) \quad \forall j, k, \ell = 1, \ldots, N.$$

With these notations, the classical formula of the homogenized coefficients is as follows:

$$q_{k\ell} = \mathcal{M}_Y\left(a_{k\ell} + a_{km}\frac{\partial\chi_\ell}{\partial y_m}\right) \quad \forall k, \ell = 1, \ldots, N.$$

(Another characterization of $q_{k\ell}$ is given in Proposition 1.5 below.) Using the above expansion, the first order corrector term is obtained in [5]. More precisely, we have the following.

Theorem 1.1. *We assume that the coefficients $a_{k\ell}$ satisfy assumptions* (1.2), *$f \in L^2(\mathbb{R}^N)$, and the solution $\chi_k \in W^{1,\infty}(Y)$, $k = 1, \ldots, N$. Then the first order corrector is defined by*

$$u_1^\varepsilon(x) = \chi_k\left(\frac{x}{\varepsilon}\right)\frac{\partial u^*}{\partial x_k}(x),$$

which means that

$$\|u^\varepsilon - u^* - \varepsilon u_1^\varepsilon\|_{H^1(\mathbb{R}^N)} \to 0 \quad \textit{as} \quad \varepsilon \to 0. \qquad \square$$

In this paper, we obtain a more general result using a different approach introduced in [11]. The basic tool of this new approach is Bloch waves ψ associated with A which we define now. Let us consider the following spectral problem parameterized by $\eta \in \mathbb{R}^N$: find $\lambda = \lambda(\eta) \in \mathbb{R}$ and $\psi = \psi(y;\eta)$ (not identically zero) such that

$$(1.15) \quad \begin{cases} A\psi(\cdot;\eta) = \lambda(\eta)\psi(\cdot;\eta) \quad \text{in} \quad \mathbb{R}^N, \quad \psi(\cdot;\eta) \text{ is } (\eta;Y)\text{-periodic, i.e.,} \\ \psi(y + 2\pi m;\eta) = e^{2\pi i m\cdot\eta}\psi(y;\eta) \quad \forall m \in \mathbb{Z}^N,\ y \in \mathbb{R}^N. \end{cases}$$

Next, we define $\phi(y;\eta) = e^{-iy\cdot\eta}\psi(y;\eta)$, and (1.15) can be rewritten in terms of ϕ as follows:

$$(1.16) \qquad A(\eta)\phi = \lambda\phi \quad \text{in} \quad \mathbb{R}^N, \quad \phi \text{ is } Y\text{-periodic.}$$

Here, the operator $A(\eta)$ is defined by

$$(1.17) \qquad A(\eta) \overset{\text{def}}{=} -\left(\frac{\partial}{\partial y_k} + i\eta_k\right)\left[a_{k\ell}(y)\left(\frac{\partial}{\partial y_\ell} + i\eta_\ell\right)\right],$$

which can be rewritten as

$$(1.18) \qquad A(\eta) = A + i\eta_k C_k + \eta_k\eta_\ell a_{k\ell}(y)$$

with

$$(1.19) \qquad C_k\phi \overset{\text{def}}{=} -a_{kj}(y)\frac{\partial\phi}{\partial y_j} - \frac{\partial}{\partial y_j}(a_{kj}(y)\phi).$$

It is clear from (1.15) that the (η, Y) periodicity condition is unaltered if we replace η by $(\eta + q)$ with $q \in \mathbb{Z}^N$, and η can therefore be confined to the *dual cell* $\eta \in Y' = [-\frac{1}{2}, \frac{1}{2}[^N$. It is well known (C. Conca, J. Planchard, and M. Vanninathan [10]) that for each $\eta \in Y'$, the above spectral problem admits a discrete sequence of eigenvalues with the following properties:

$$\begin{cases} 0 \le \lambda_1(\eta) \le \cdots \le \lambda_m(\eta) \le \cdots \to \infty \\ \forall m \ge 1,\ \lambda_m(\eta) \text{ is a Lipschitz function of } \eta \in Y'. \end{cases}$$

Besides, the corresponding eigenfunctions denoted by $\psi_m(\cdot;\eta)$ and $\phi_m(\cdot;\eta)$ form orthonormal bases in the spaces of all $L^2_{loc}(\mathbb{R}^N)$-functions which are $(\eta;Y)$-periodic and Y-periodic, respectively; these spaces are denoted by $L^2_\#(\eta;Y)$ and $L^2_\#(Y)$. It is worthwhile to remark that these eigenfunctions in fact belong to the spaces $H^1_\#(\eta;Y)$ and $H^1_\#(Y)$, respectively, where

$$H^1_\#(\eta;Y) = \left\{\psi \in L^2_\#(\eta;Y) \;\middle|\; \frac{\partial\psi}{\partial y_k} \in L^2_\#(\eta;Y) \quad \forall k = 1,\dots,N\right\},$$

$$H^1_\#(Y) = \left\{\phi \in L^2_\#(Y) \;\middle|\; \frac{\partial\phi}{\partial y_k} \in L^2_\#(Y) \quad \forall k = 1,\dots,N\right\}.$$

The functions $\psi_m(\cdot;\eta)$ and $\phi_m(\cdot;\eta)$ (referred to as *Bloch waves*) introduced above enable us to describe the spectral resolution of A (an unbounded self-adjoint operator in $L^2(\mathbb{R}^N)$) in the orthogonal basis $\{e^{iy\cdot\eta}\phi_m(y;\eta)|m \geq 1, \eta \in Y'\}$. More precisely, we have the following.

THEOREM 1.2. *Let $g \in L^2(\mathbb{R}^N)$. The mth Bloch coefficient of g is defined as follows:*

$$(B_m g)(\eta) = \int_{\mathbb{R}^N} g(y)e^{-iy\cdot\eta}\bar{\phi}_m(y;\eta)dy \quad \forall m \geq 1,\ \eta \in Y'.$$

Then the following inverse formula holds:

$$g(y) = \int_{Y'} \sum_{m=1}^{\infty} (B_m g)(\eta)e^{iy\cdot\eta}\phi_m(y;\eta)d\eta.$$

Further, we have Parseval's identity:

$$\int_{\mathbb{R}^N} |g(y)|^2 dy = \int_{Y'} \sum_{m=1}^{\infty} |(B_m g)(\eta)|^2 d\eta.$$

Finally, for all g in the domain of A, we have

$$Ag(y) = \int_{Y'} \sum_{m=1}^{\infty} \lambda_m(\eta)(B_m g)(\eta)e^{iy\cdot\eta}\phi_m(y;\eta)d\eta. \qquad \square$$

To obtain the spectral resolution of A^ε in an analogous manner, let us introduce Bloch waves at the ε-scale:

$$\lambda_m^\varepsilon(\xi) = \varepsilon^{-2}\lambda_m(\eta), \quad \phi_m^\varepsilon(x;\xi) = \phi_m(y;\eta), \quad \psi_m^\varepsilon(x;\xi) = \psi_m(y;\eta),$$

where the variables (x,ξ) and (y,η) are related by $y = \frac{x}{\varepsilon}$ and $\eta = \varepsilon\xi$. Observe that $\phi_m^\varepsilon(x;\xi)$ is εY-periodic (in x) and $\varepsilon^{-1}Y'$-periodic with respect to ξ. In the same manner, $\psi_m^\varepsilon(\cdot;\xi)$ is $(\varepsilon\xi;\varepsilon Y)$-periodic because of the relation $\psi_m^\varepsilon(x;\xi) = e^{ix\cdot\xi}\phi_m^\varepsilon(x;\xi)$. Note that the dual cell at ε-scale is $\varepsilon^{-1}Y'$ and hence we take ξ to vary in $\varepsilon^{-1}Y'$ in what follows. With these notations, we have the following result analogous to Theorem 1.2.

THEOREM 1.3. *Let $g \in L^2(\mathbb{R}^N)$. The mth Bloch coefficient of g at the ε-scale is defined as follows:*

$$(B_m^\varepsilon g)(\xi) = \int_{\mathbb{R}^N} g(x)e^{-ix\cdot\xi}\bar{\phi}_m^\varepsilon(x;\xi)dx \quad \forall m \geq 1,\ \xi \in \varepsilon^{-1}Y'.$$

Then the following inverse formula and Parseval's identity hold:

$$g(x) = \int_{\varepsilon^{-1}Y'} \sum_{m=1}^{\infty} (B_m^\varepsilon g)(\xi)e^{ix\cdot\xi}\phi_m^\varepsilon(x;\xi)d\xi,$$

$$\int_{\mathbb{R}^N} |g(x)|^2 dx = \int_{\varepsilon^{-1}Y'} \sum_{m=1}^{\infty} |(B_m^\varepsilon g)(\xi)|^2 d\xi.$$

Finally, for all g in the domain of A^ε, we get

$$A^\varepsilon g(x) = \int_{\varepsilon^{-1}Y'} \sum_{m=1}^{\infty} \lambda_m^\varepsilon(\xi)(B_m^\varepsilon g)(\xi)e^{ix\cdot\xi}\phi_m^\varepsilon(x;\xi)d\xi. \qquad \square$$

Using the above theorem, the classical homogenization result was deduced in [11]. Let us recall the main steps. The first one consists of considering a sequence $u^\varepsilon \in H^1(\mathbb{R}^N)$ satisfying (1.9). We can express the equation $A^\varepsilon u^\varepsilon = f$ in $\mathbb{R}^N$ in the equivalent form

$$\lambda_m^\varepsilon(\xi)(B_m^\varepsilon u^\varepsilon)(\xi) = (B_m^\varepsilon f)(\xi) \quad \forall m \geq 1,\ \xi \in \varepsilon^{-1}Y'. \tag{1.20}$$

In the homogenization process, one can neglect all the relations for $m \geq 2$. More precisely, it is proved in [11] that the following result holds.

PROPOSITION 1.4. *Let*

$$v^\varepsilon(x) = \int_{\varepsilon^{-1}Y'} \sum_{m=2}^{\infty} (B_m^\varepsilon u^\varepsilon)(\xi) e^{ix\cdot\xi} \phi_m^\varepsilon(x;\xi) d\xi. \tag{1.21}$$

Then $\|v^\varepsilon\|_{L^2(\mathbb{R}^N)} \leq c\varepsilon$. □

Thus we can concentrate our attention only on the relation corresponding to the first Bloch wave:

$$\lambda_1^\varepsilon(\xi)(B_1^\varepsilon u^\varepsilon)(\xi) = (B_1^\varepsilon f)(\xi) \quad \forall \xi \in \varepsilon^{-1}Y'. \tag{1.22}$$

The homogenized equation in the Fourier space

$$q_{k\ell}\xi_k\xi_\ell \widehat{u*}(\xi) = \widehat{f}(\xi) \quad \forall \xi \in \mathbb{R}^N \tag{1.23}$$

is obtained from (1.22) by passing to the limit as $\varepsilon \to 0$. Here, the symbol $\widehat{\ }$ stands for the classical Fourier transformation

$$\widehat{f}(\xi) = \frac{1}{(2\pi)^{N/2}} \int_{\mathbb{R}^N} f(x) e^{-ix\cdot\xi} dx.$$

To this end, the following results were established and applied in [11].

PROPOSITION 1.5. *We assume that* $a_{k\ell}$ *satisfies* (1.2). *Then there exists* $\delta > 0$ *such that the first eigenvalue* $\lambda_1(\eta)$ *is an analytic function on* $B_\delta \overset{\text{def}}{=} \{\eta \mid |\eta| < \delta\}$, *and there is a choice of the first eigenvector* $\phi_1(y;\eta)$ *satisfying*

$$\begin{cases} \eta \to \phi_1(\cdot;\eta) \in H^1_\#(Y) \text{ is analytic on } B_\delta, \\ \phi_1(y;0) = p^{(0)} \ (= |Y|^{-1/2} = \frac{1}{(2\pi)^{N/2}}). \end{cases}$$

Moreover, we have the relations

$$\lambda_1(0) = 0, \quad D_k\lambda_1(0) = \frac{\partial \lambda_1}{\partial \eta_k}(0) = 0 \quad \forall k = 1, \ldots, N,$$

$$\frac{1}{2} D^2_{k\ell}\lambda_1(0) = \frac{1}{2}\frac{\partial^2 \lambda_1}{\partial \eta_k \partial \eta_\ell}(0) = q_{k\ell} \quad \forall k, \ell = 1, \ldots, N,$$

and there exist constants c *and* $\widetilde{c}$ *such that*

$$c|\eta|^2 \leq \lambda_1(\eta) \leq \widetilde{c}|\eta|^2 \quad \forall \eta \in Y', \tag{1.24}$$

$$0 < \lambda_2^{(N)} \leq \lambda_m(\eta) \quad \forall m \geq 2,\ \eta \in Y', \tag{1.25}$$

where $\lambda_2^{(N)}$ *is the second eigenvalue of the spectral problem for* A *in the cell* Y *with Neumann boundary conditions on* ∂Y. □

Apart from the above result of regularity on the Bloch spectrum, we need to prove that the first Bloch transform is an approximation to the Fourier transform. This result is naturally expected from the fact that $\phi_1^\varepsilon(x;\xi) \to (2\pi)^{-N/2}$, as $\varepsilon \to 0$, $\forall \xi \in \mathbb{R}^N$.

PROPOSITION 1.6. *Let g^ε and g be in $L^2(\mathbb{R}^N)$. Then*

(i) *if $g^\varepsilon \rightharpoonup g$ weakly in $L^2(\mathbb{R}_x^N)$, then $\chi_{\varepsilon^{-1}Y'} B_1^\varepsilon g^\varepsilon \rightharpoonup \widehat{g}$ weakly in $L^2_{loc}(\mathbb{R}_\xi^N)$ provided there is a fixed compact set K such that* $\operatorname{supp}(g^\varepsilon) \subset K \, \forall \varepsilon$;

(ii) *if $g^\varepsilon \to g$ in $L^2(\mathbb{R}_x^N)$, then $\chi_{\varepsilon^{-1}Y'} B_1^\varepsilon g^\varepsilon \to \widehat{g}$ in $L^2_{loc}(\mathbb{R}_\xi^N)$.* □

These results easily lead us to the following homogenization theorem in $\mathbb{R}^N$.

THEOREM 1.7. *We consider a sequence u^ε satisfying* (1.9). *Then*

$$a_{k\ell}^\varepsilon \frac{\partial u^\varepsilon}{\partial x_\ell} \rightharpoonup q_{k\ell} \frac{\partial u^*}{\partial x_\ell} \quad \text{in} \quad L^2(\mathbb{R}^N) \quad \forall k = 1, \dots, N.$$

In particular, u^* satisfies $A^* u^* = f$ in $\mathbb{R}^N$. □

Once the homogenization result in $\mathbb{R}^N$ is established, it is an easy matter to deduce the corresponding result in a bounded domain Ω by localization techniques using a cut-off function $\phi \in \mathcal{D}(\Omega)$ (see [11]).

1.2. Presentation of new results: The Bloch approximation. Let us consider the sequence u^ε satisfying hypotheses (1.9). The *Bloch approximation* of u^ε is defined by the following formula:

$$\theta^\varepsilon(x) \overset{\text{def}}{=} \int_{\varepsilon^{-1}Y'} \widehat{u^*}(\xi) e^{ix\cdot\xi} \phi_1^\varepsilon(x;\xi) d\xi, \quad x \in \mathbb{R}^N. \tag{1.26}$$

First of all, let us remark that this object is not difficult to be computed in principle. Our goal throughout this paper is to study properties of this function and particularly its relations with various correctors terms. It is worth noticing that θ^ε is defined only in terms of the first Bloch mode ϕ_1^ε. We will see in section 3 that higher Bloch modes ϕ_m^ε, $m \geq 2$, do not contribute at all in the analysis of the correctors of first and second order in the energy norm. (It will be interesting to know whether these higher order modes play a part in the analysis of correctors in stronger norms $H^2, \dots,$ etc. For H^2-estimates, we refer to our work in [12].) Thus we are motivated to introduce the projection onto the first Bloch mode: for all $g \in L^2(\mathbb{R}^N)$, we define

$$P_1^\varepsilon g(x) = \int_{\varepsilon^{-1}Y'} B_1^\varepsilon g(\xi) e^{ix\cdot\xi} \phi_1^\varepsilon(x;\xi) d\xi, \quad x \in \mathbb{R}^N. \tag{1.27}$$

We note by the item (ii) of Proposition 1.6 that the Fourier transform $\widehat{u*}$ is an approximation of $B_1^\varepsilon u^\varepsilon$. Therefore, heuristically speaking, the Bloch approximation θ^ε is close to $P_1^\varepsilon u^\varepsilon$ and hence to u^ε. With these notations, we will prove the following theorem.

THEOREM 1.8. *Assume that the coefficients $a_{k\ell}$ satisfy* (1.2). *Let u^ε be the sequence introduced in* (1.9). *Then if $f \in L^2(\mathbb{R}^N)$, we have*

$$(u^\varepsilon - \theta^\varepsilon) \to 0 \quad \text{in} \quad H^1(\mathbb{R}^N). \tag{1.28}$$

Furthermore, we have the estimate

$$|u^\varepsilon - \theta^\varepsilon|_{H^1(\mathbb{R}^N)} \leq c\varepsilon \|f\|_{L^2(\mathbb{R}^N)}. \tag{1.29}$$ □

It is worth remarking that even though error estimates of the type (1.29) are sometimes found in the literature, they are usually obtained using the maximum principle with more regularity hypotheses on $a_{k\ell}$ and f. Here, we obtain these natural estimates under optimal hypotheses.

Thanks to the above result, we are reduced to expanding θ^ε in terms of ε in order to be able to compare it with the classical correctors for u^ε. To fulfill this task, it is clear from the definition of θ^ε that it is necessary to obtain asymptotic expansions of the first eigenvalue $\lambda_1^\varepsilon(\xi)$ and the first Bloch mode $\phi_1^\varepsilon(\cdot;\xi)$. (In addition, for our purposes below, we need an asymptotic expansion of the first Bloch transform $B_1^\varepsilon g(\xi)$ for which we refer the reader to section 5. These results strengthen earlier results, particularly those of Proposition 1.6.) We now state results in this direction, and their proofs will be taken up in the following sections along with other auxiliary results. First, we introduce some test functions $\chi_{k\ell}$, $\chi_{k\ell m}$, $\chi_{k\ell mn}$ defined by the following cell problems (observe that the first ones are nothing but the functions already introduced in (1.14)):

$$(1.30)\qquad \begin{cases} A\chi_{k\ell} = (a_{k\ell} - q_{k\ell}) - \dfrac{1}{2}\left(C_k\chi_\ell + C_\ell\chi_k\right) \quad \text{in} \quad \mathbb{R}^N, \\ \chi_{k\ell} \in H^1_\#(Y), \quad \mathcal{M}_Y(\chi_{k\ell}) = 0. \end{cases}$$

$$(1.31)\qquad \begin{cases} A\chi_{k\ell m} = \dfrac{1}{3}\Big[(a_{k\ell} - q_{k\ell})\chi_m + (a_{\ell m} - q_{\ell m})\chi_k + (a_{mk} - q_{mk})\chi_\ell \\ \qquad\qquad -C_k\chi_{\ell m} - C_\ell\chi_{mk} - C_m\chi_{k\ell}\Big] \quad \text{in} \quad \mathbb{R}^N, \\ \chi_{k\ell m} \in H^1_\#(Y), \quad \mathcal{M}_Y(\chi_{k\ell m}) = 0. \end{cases}$$

$$(1.32)\qquad \begin{cases} A\chi_{k\ell mn} = \dfrac{1}{4!}D^4_{k\ell mn}\lambda_1(0) - \dfrac{1}{4}\left(C_n\chi_{k\ell m} + C_k\chi_{\ell mn} + C_\ell\chi_{mnk} + C_m\chi_{nk\ell}\right) \\ \qquad + \dfrac{1}{3!}\Big[(a_{k\ell} - q_{k\ell})\chi_{mn} + (a_{\ell m} - q_{\ell m})\chi_{kn} + (a_{km} - q_{km})\chi_{\ell n} \\ \qquad + (a_{mn} - q_{mn})\chi_{k\ell} + (a_{\ell n} - q_{\ell n})\chi_{km} + (a_{kn} - q_{kn})\chi_{\ell m}\Big] \text{ in } \mathbb{R}^N, \\ \chi_{k\ell mn} \in H^1_\#(Y), \quad \mathcal{M}_Y(\chi_{k\ell mn}) = 0. \qquad \square \end{cases}$$

Proposition 1.9. *All odd order derivatives of λ_1 at $\eta = 0$ vanish, i.e.,*

$$D^\beta\lambda_1(0) = 0 \quad \forall\beta \in \mathbb{Z}^N_+,\ |\beta| \text{ odd.}$$

All even order derivatives of λ_1 at $\eta = 0$ can be calculated in a systematic fashion. For instance, the fourth order derivatives have the following expressions: for all $k, \ell, m, n = 1, \ldots, N$

$$\frac{1}{4!}D^4_{k\ell mn}\lambda_1(0) = \frac{1}{4}\frac{1}{|Y|}\int_Y \left\{C_n\chi_{k\ell m} + C_k\chi_{\ell mn} + C_\ell\chi_{mnk} + C_m\chi_{nk\ell}\right\} dy$$

$$- \frac{1}{3!}\frac{1}{|Y|}\int_Y \Big\{ (a_{k\ell} - q_{k\ell})\chi_{mn} + (a_{\ell m} - q_{\ell m})\chi_{nk} + (a_{mn} - q_{mn})\chi_{k\ell}$$

$$+ \ (a_{nk} - q_{nk})\chi_{\ell m} + (a_{km} - q_{km})\chi_{\ell n} + (a_{\ell n} - q_{\ell n})\chi_{km} \Big\}\, dy. \qquad \square$$

Various derivatives of ϕ_1 at $\eta = 0$ can also be calculated in a systematic fashion.

PROPOSITION 1.10. *We have the following expressions:*

$$D_k\phi_1(y;0) = ip^{(0)}\chi_k(y),$$

$$\frac{1}{2!}D^2_{k\ell}\phi_1(y;0) = -\, p^{(0)}\chi_{k\ell}(y) - \beta^{(2)}_{k\ell}p^{(0)},$$

$$\frac{1}{3!}D^3_{k\ell m}\phi_1(y;0) = -\, ip^{(0)}\chi_{k\ell m}(y) - \frac{i}{3}\left(\beta^{(2)}_{k\ell}\chi_m(y) + \beta^{(2)}_{\ell m}\chi_k(y) + \beta^{(2)}_{mk}\chi_\ell(y)\right)p^{(0)},$$

$$\frac{1}{4!}D^4_{k\ell mn}\phi_1(y;0) = p^{(0)}\chi_{k\ell mn}(y) - \frac{1}{3!}\Big(\beta^{(2)}_{k\ell}\chi_{mn}(y) + \beta^{(2)}_{\ell m}\chi_{nk}(y) + \beta^{(2)}_{mn}\chi_{k\ell}$$

$$+ \beta^{(2)}_{nk}\chi_{\ell m}(y) + \beta^{(2)}_{km}\chi_{n\ell}(y) + \beta^{(2)}_{\ell n}\chi_{km}(y)\Big)p^{(0)} + \beta^{(4)}_{k\ell mn}p^{(0)}$$

with

$$\beta^{(2)}_{k\ell} = \frac{1}{2!}\frac{1}{|Y|}\int_Y \chi_\ell\chi_k dy,$$

$$\beta^{(4)}_{k\ell mn} = \frac{1}{|Y|}\int_Y \frac{1}{4}\left[\chi_{\ell mn}\chi_k + \chi_{kmn}\chi_\ell + \chi_{n\ell k}\chi_m + \chi_{k\ell n}\chi_n\right]dy$$

$$- \frac{1}{|Y|}\int_Y \frac{1}{6}\left[\chi_{\ell m}\chi_{kn} + \chi_{km}\chi_{n\ell} + \chi_{\ell k}\chi_{nm}\right]dy$$

$$+ \frac{1}{|Y|}\frac{1}{2}\left(\beta^{(2)}_{k\ell}\beta^{(2)}_{mn} + \beta^{(2)}_{km}\beta^{(2)}_{n\ell} + \beta^{(2)}_{kn}\beta^{(2)}_{m\ell}\right). \qquad \square$$

We note that all odd order derivatives of ϕ_1 at $\eta = 0$ are purely imaginary and all even order derivatives are real.

Since $\phi_1(\cdot;\eta)$ is proved to be analytic for $|\eta| \le \delta$, we can expand it and thus give rise to an asymptotic expansion of θ^ε which is as follows:

$$(1.33)\quad \theta^\varepsilon(x) = u^*(x) + \varepsilon\chi_k\left(\frac{x}{\varepsilon}\right)\frac{\partial u^*}{\partial x_k}(x) - \varepsilon^2\left(\chi_{k\ell}\left(\frac{x}{\varepsilon}\right) + \beta^{(2)}_{k\ell}\right)\frac{\partial^2 u^*}{\partial x_k \partial x_\ell}(x) + \cdots.$$

This can be rigorously proved. Our next result is a sample where we specify the precise hypotheses needed to justify the above expansion up to three terms.

THEOREM 1.11. *Assume that the hypotheses of Theorem* 1.8 *hold.*

(i) *If* $u^* \in H^1(\mathbb{R}^N)$, *then*

$$\|\theta^\varepsilon - u^*\|_{L^2(\mathbb{R}^N)} \le c\varepsilon\|u^*\|_{H^1(\mathbb{R}^N)}.$$

(ii) *If* $f \in L^2(\mathbb{R}^N)$ *and* $\chi_k \in W^{1,\infty}_{\#}(Y)$, *where* χ_k *is the solution of* (1.13) *and* $\chi^\varepsilon_k(x) = \chi_k\left(\frac{x}{\varepsilon}\right)$, *then we have*

$$\left\|\theta^\varepsilon - u^* - \varepsilon\chi^\varepsilon_k\frac{\partial u^*}{\partial x_k}\right\|_{H^1(\mathbb{R}^N)} \le c\varepsilon\|f\|_{L^2(\mathbb{R}^N)}.$$

(iii) *If* $f \in H^1(\mathbb{R}^N)$ *and* χ_k, $\chi_{k\ell} \in W^{1,\infty}_{\#}(Y)$, *where* $\chi_{k\ell}$ *is the solution of* (1.30), $\beta^{(2)}_{k\ell}$ *are constants defined in Proposition* 1.10, *and* $\chi^\varepsilon_{k\ell}(x) = \chi_{k\ell}\left(\frac{x}{\varepsilon}\right)$, *then*

$$\left\| \theta^\varepsilon - u^* - \varepsilon \chi^\varepsilon_k \frac{\partial u^*}{\partial x_k} + \varepsilon^2 \left(\chi^\varepsilon_{k\ell} + \beta^{(2)}_{k\ell}\right) \frac{\partial^2 u^*}{\partial x_k \partial x_\ell} \right\|_{H^1(\mathbb{R}^N)} \leq c\varepsilon^2 \|f\|_{H^1(\mathbb{R}^N)}. \qquad \square$$

It is important to note that these above expansions are of Taylor type owing to the analyticity of $\lambda_1(\eta)$ and $\phi_1(\cdot;\eta)$. This is the main difference between this approach and the classical one found in [5], where the expansion has a multiscale structure.

Concerning the hypotheses on the smoothness of functions χ_k and $\chi_{k\ell}$ in statements (ii) and (iii), it is worth mentioning from regularity theory of elliptic boundary value problems that $W^{1,\infty}$-estimates are hard to come by. This is why they are usually assumed in homogenization theory. However, several numerical studies with simple fibers show that these assumptions are valid. Thus they are reasonable hypotheses to work with as far as certain applications are concerned.

The expansions of $\lambda_1(\eta)$, $\phi^\varepsilon_1(\cdot;\eta)$, and $B^\varepsilon_1 g(\xi)$ obtained in Propositions 1.5 and 1.9 and Propositions 5.1, 5.2, and 5.3 below have further interesting consequences which will be developed in a forthcoming paper. For the time being, we will be content with a few remarks. Since higher order modes can be neglected, the first eigenvalue $\lambda_1(\eta)$ along with the first eigenvector $\phi_1(\cdot;\eta)$ represent the periodic medium under consideration. Their contributions occur somewhat separately without interaction at the levels of homogenized equation and correctors. More precisely, the first eigenvalue $\lambda_1(\eta)$ contributes at various levels through its derivatives at $\eta = 0$. The first eigenvector $\phi_1(\cdot;\eta)$ and its first derivatives contribute through the first Bloch transform $B^\varepsilon_1 g(\xi)$ and its expansion described in Propositions 5.2 and 5.3.

In the homogenized equation, for instance, we see the product of the second order derivatives of $\lambda_1(\eta)$ at $\eta = 0$ with the 0th order term of $B^\varepsilon_1 g(\xi)$, namely, $\widehat{g}(\xi)$. We see a similar structure in the correctors, too. There seem to be situations where both interact and contribute jointly in a manner different from the above. One example of such a situation is the study of the propagation of waves in a periodic medium. It appears that the homogenized medium is not good enough to provide an approximation to the propagation for large times because of the appearance of dispersion effects shown numerically in F. Santosa and W. W. Symes [20]. We feel that this is an appropriate place to highlight the improvements achieved in this work with respect to [20]. Apart from the mathematical rigor, the main point is that the third order derivatives of $\lambda_1(\eta)$ at $\eta = 0$ are shown to be zero even in the multidimensional case. (In fact all odd order derivatives vanish.) Moreover, our arguments are more general compared with the one-dimensional case covered in [20]. This will have consequences in the propagation of waves in periodic media. We plan to cover these aspects in a future publication.

We conclude this introduction by saying how the rest of this paper is organized. Section 2 is devoted to certain fundamental lemmas which are indispensable. As an immediate application, we prove in section 3 that the higher order Bloch modes are negligible. Taylor expansions for λ_1 and ϕ_1 are obtained in section 4 which proves Propositions 1.9 and 1.10. Section 5 is devoted to the description of the asymptotic behavior of the first Bloch transform B^ε_1 whose definition is given in Theorem 1.2. Finally, in section 6, we present the proofs of the main results, namely, Theorems 1.8 and 1.11.

2. Fundamental lemmas. In this section, we prove a series of results which generalize Parseval's identity stated in Theorem 1.3. These estimates will be useful later for the analysis of correctors. The following two lemmas are easily seen to be generalizations of well-known classical results for $-\Delta$.

LEMMA 2.1. *For all $g \in H^1(\mathbb{R}^{\mathbb{N}})$, we have*

$$c_1 |g|^2_{H^1(\mathbb{R}^N)} \le \int_{\varepsilon^{-1}Y'} \sum_{m=1}^{\infty} \lambda_m^\varepsilon(\xi) |B_m^\varepsilon g(\xi)|^2 d\xi \le c_2 |g|^2_{H^1(\mathbb{R}^N)},$$

where c_1 and c_2 are constants independent of ε and g.

Proof. First of all, by uniform ellipticity of A^ε, we have

$$\alpha \int_{\mathbb{R}^N} |\nabla g|^2 dx \le \int_{\mathbb{R}^N} A^\varepsilon g \bar{g} dx \le \beta \int_{\mathbb{R}^N} |\nabla g|^2 dx.$$

We can rewrite the middle term by applying the Plancherel identity:

$$\int_{\mathbb{R}^N} g(x)\overline{h(x)} dx = \int_{\varepsilon^{-1}Y'} \sum_{m=1}^{\infty} B_m^\varepsilon g(\xi) \overline{B_m^\varepsilon h(\xi)} d\xi \quad \forall g, h \in L^2(\mathbb{R}^N). \tag{2.1}$$

Indeed, using the spectral resolution of A^ε, we get

$$\int_{\mathbb{R}^N} A^\varepsilon g \bar{g} dx = \int_{\varepsilon^{-1}Y'} \sum_{m=1}^{\infty} \lambda_m^\varepsilon(\xi) |B_m^\varepsilon g(\xi)|^2 d\xi.$$

This completes the proof. □

By using the duality between $H^1(\mathbb{R}^N)$ and $H^{-1}(\mathbb{R}^N)$, we deduce the following.

LEMMA 2.2. *For all $g \in H^{-1}(\mathbb{R}^N)$, there exist c_1 and c_2 which are independent of ε and g, such that*

$$c_1 \|g\|^2_{H^{-1}(\mathbb{R}^N)} \le \int_{\varepsilon^{-1}Y'} \sum_{m=1}^{\infty} \frac{1}{1+\lambda_m^\varepsilon(\xi)} |B_m^\varepsilon g(\xi)|^2 d\xi \le c_2 \|g\|^2_{H^{-1}(\mathbb{R}^N)}.$$

Proof. It is well known that $(A^\varepsilon + I)\colon H^1(\mathbb{R}^N) \to H^{-1}(\mathbb{R}^N)$ is an isomorphism. For every $g \in H^{-1}(\mathbb{R}^N)$ there exists a unique solution $u \in H^1(\mathbb{R}^N)$ of $A^\varepsilon u + u = g$ in $\mathbb{R}^N$. We can express the previous equation in the equivalent form

$$(\lambda_m^\varepsilon(\xi) + 1) B_m^\varepsilon u(\xi) = B_m^\varepsilon g(\xi) \quad \forall m \ge 1,\ \xi \in \varepsilon^{-1}Y'.$$

Therefore, an application of the Cauchy–Schwarz inequality yields

$$\begin{aligned}
\langle g, v\rangle &= \int_{\varepsilon^{-1}Y'} \sum_{m=1}^{\infty} (\lambda_m^\varepsilon(\xi)+1) B_m^\varepsilon u B_m^\varepsilon v d\xi \\
&\le \left(\int_{\varepsilon^{-1}Y'} \sum_{m=1}^{\infty} (\lambda_m^\varepsilon(\xi)+1) |B_m^\varepsilon u|^2 d\xi \right)^{1/2} \left(\int_{\varepsilon^{-1}Y'} \sum_{m=1}^{\infty} (\lambda_m^\varepsilon(\xi)+1) |B_m^\varepsilon v|^2 d\xi \right)^{1/2} \\
&\le \left(\int_{\varepsilon^{-1}Y'} \sum_{m=1}^{\infty} \frac{|B_m^\varepsilon g|^2}{(\lambda_m^\varepsilon(\xi)+1)} d\xi \right)^{1/2} \left(\int_{\varepsilon^{-1}Y'} \sum_{m=1}^{\infty} (\lambda_m^\varepsilon(\xi)+1) |B_m^\varepsilon v|^2 d\xi \right)^{1/2}
\end{aligned}$$

for all $v \in H^1(\mathbb{R}^N)$, $g \in H^{-1}(\mathbb{R}^N)$. Here, $\langle \cdot, \cdot \rangle$ denotes the $H^1(\mathbb{R}^N)$ and $H^{-1}(\mathbb{R}^N)$ duality pairing. By virtue of Lemma 2.1 and Parseval's identity, the second term in the right-hand side is equivalent to the H^1-norm of v. Thus we deduce the existence of a constant c_1 such that

$$c_1 \|g\|^2_{H^{-1}(\mathbb{R}^N)} \leq \int_{\varepsilon^{-1}Y'} \sum_{m=1}^{\infty} \frac{1}{1+\lambda_m^\varepsilon(\xi)} |B_m^\varepsilon g(\xi)|^2 d\xi,$$

which is the lower estimate in Lemma 2.2. To prove the upper estimate is enough to use the continuity of the solution $u \in H^1(\mathbb{R}^N)$ with respect to the right-hand side $g \in H^{-1}(\mathbb{R}^N)$. □

In our next lemma, we consider $g^\varepsilon = g^\varepsilon(\xi)$ a measurable function defined on $\varepsilon^{-1}Y'$, and another function $\rho = \rho(y;\eta)$ measurable with respect to $(y;\eta)$ and Y-periodic in y. We then introduce

$$G^\varepsilon(x) = \int_{\varepsilon^{-1}Y'} g^\varepsilon(\xi) e^{ix\cdot\xi} \rho\left(\frac{x}{\varepsilon};\varepsilon\xi\right) d\xi, \quad x \in \mathbb{R}^N. \tag{2.2}$$

The following result estimates its $L^2(\mathbb{R}^N)$ and $H^1(\mathbb{R}^N)$ norms.

LEMMA 2.3. *We assume $g^\varepsilon \in L^2(\varepsilon^{-1}Y')$ and $\rho \in L^\infty(Y'; H^1_\#(Y))$. Then we have*

$$\|G^\varepsilon\|^2_{L^2(\mathbb{R}^N)} = \int_{\varepsilon^{-1}Y'} |g^\varepsilon(\xi)|^2 \|\rho(\cdot;\varepsilon\xi)\|^2_{L^2(Y)} d\xi,$$

$$|G^\varepsilon|^2_{H^1(\mathbb{R}^N)} = \int_{\varepsilon^{-1}Y'} |g^\varepsilon(\xi)|^2 \|i\xi\rho(\cdot;\varepsilon\xi) + \varepsilon^{-1}\nabla_y\rho(\cdot;\varepsilon\xi)\|^2_{L^2(Y)^N} d\xi.$$

Proof. We expand $\rho(y;\eta)$ as a function of y in the orthonormal basis $\{\phi_m(y;\eta)\}_{m=1}^\infty$ where η is a parameter:

$$\rho(y;\eta) = \sum_{m=1}^{\infty} a_m(\eta)\phi_m(y;\eta).$$

Introducing this expression in (2.2), we get

$$G^\varepsilon(x) = \int_{\varepsilon^{-1}Y'} g^\varepsilon(\xi) \sum_{m=1}^{\infty} a_m(\varepsilon\xi) e^{ix\cdot\xi} \phi_m^\varepsilon(x;\xi) d\xi.$$

Applying Parseval's identity of Theorem 1.3, we get

$$\|G^\varepsilon\|^2_{L^2(\mathbb{R}^N)} = \int_{\varepsilon^{-1}Y'} |g^\varepsilon(\xi)|^2 \sum_{m=1}^{\infty} |a_m(\varepsilon\xi)|^2 d\xi.$$

This completes the proof of the first part of the lemma if we use Parseval's identity in $L^2(Y)$:

$$\|\rho(\cdot;\eta)\|^2_{L^2(Y)} = \sum_{m=1}^{\infty} |a_m(\eta)|^2 \quad \forall \eta \in Y'. \tag{2.3}$$

For the second part of the lemma, we formally differentiate $G^\varepsilon(x)$ with respect to x. We obtain

$$\nabla_x G^\varepsilon(x) = \int_{\varepsilon^{-1}Y'} g^\varepsilon(\xi) e^{ix\cdot\xi} \left(i\xi\rho\left(\frac{x}{\varepsilon};\varepsilon\xi\right) + \varepsilon^{-1}\nabla_y\rho\left(\frac{x}{\varepsilon};\varepsilon\xi\right)\right) d\xi.$$

We remark that the above integral is of the same type as the one analyzed in the first part. This completes the proof. □

The next lemma presents H^1-estimates on the Bloch modes.

LEMMA 2.4. *We suppose that the coefficients $a_{k\ell}$ satisfy* (1.2). *Then there exists a constant c depending on $\|a_{k\ell}\|_{L^\infty(Y)}$ such that*

$$\left\|\frac{\partial\phi_m}{\partial y_k}(\cdot;\eta)\right\|_{L^2(Y)} \leq c_1\lambda_m(\eta)^{1/2} \quad \forall\eta\in Y',\ m\geq 1,\ k=1,\dots,N. \tag{2.4}$$ □

To prove this, let us introduce the bilinear forms associated with the operators $A(\eta)$ and A, respectively.

$$a(\eta;\phi,\psi) = \int_Y a_{k\ell}(y)\left(\frac{\partial\phi}{\partial y_\ell}+i\eta_\ell\phi\right)\overline{\left(\frac{\partial\psi}{\partial y_k}+i\eta_k\psi\right)}dy,$$

$$a(\phi,\psi) = \int_Y a_{k\ell}(y)\frac{\partial\phi}{\partial y_\ell}\overline{\frac{\partial\psi}{\partial y_k}}dy.$$

The basic estimates on them are obtained in [10, p. 190]: There exist constants c, $\widetilde{c}$ which are independent of $\eta\in Y'$ such that for all $\phi\in H^1_\#(Y)$,

$$c\ (\|\nabla\phi\|^2_{L^2(Y)^N}+|\eta|^2\|\phi\|^2_{L^2(Y)}) \leq a(\eta;\phi,\phi) \leq \widetilde{c}\ (\|\nabla\phi\|^2_{L^2(Y)^N}+|\eta|^2\|\phi\|^2_{L^2(Y)}), \tag{2.5}$$

$$c\ \|\nabla\phi\|^2_{L^2(Y)^N} \leq\ a(\phi,\phi)\ \leq \widetilde{c}\ \|\nabla\phi\|^2_{L^2(Y)^N}. \tag{2.6}$$

Proof of Lemma 2.4. For simplicity, we denote $\phi_m(\cdot;\eta)$ by $\phi_m(\eta)$. We recall that it satisfies

$$a(\eta;\phi_m(\eta),\psi) = \lambda_m(\eta)(\phi_m(\eta),\psi) \quad \forall\psi\in H^1_\#(Y). \tag{2.7}$$

To deduce (2.4), it is enough to take $\psi=\phi_m(\eta)$ and use (2.5). □

Our next result concerns the estimation of expressions which are inverse to (2.2). We define

$$J^\varepsilon g(\xi) = \int_{\mathbb{R}^N} g(x)e^{-ix\cdot\xi}\rho\left(\frac{x}{\varepsilon};\varepsilon\xi\right)dx \quad \text{for} \quad \xi\in\varepsilon^{-1}Y', \tag{2.8}$$

where $g=g(x)$ is a measurable function defined on $\mathbb{R}^N$ and $\rho=\rho(y;\eta)$ is a measurable function defined on $Y\times Y'$. We assume that ρ is Y-periodic in y. The required hypotheses on these functions will depend on the estimate deduced on $J^\varepsilon g$. This is illustrated in the results that follow which are analogous to classical estimates on the Fourier transform.

LEMMA 2.5.

(i) *If $g\in L^2(\mathbb{R}^N)$ and $\rho\in L^\infty(Y';L^2_\#(Y))$, then we have*

$$\|J^\varepsilon g\|_{L^2(\varepsilon^{-1}Y')} \leq \|g\|_{L^2(\mathbb{R}^N)}\|\rho\|_{L^\infty(Y';L^2_\#(Y))}.$$

(ii) *If $g\in H^1(\mathbb{R}^N)$ and $\rho\in L^\infty(Y';H^1_\#(Y))$, then we have*

$$\|(1+|\xi|^2)^{1/2}J^\varepsilon g(\xi)\|_{L^2(\varepsilon^{-1}Y')} \leq c\left\{\|\nabla g\|_{L^2(\mathbb{R}^N)}\|\rho\|_{L^\infty(Y';L^2(Y))}\right.$$
$$\left.+\ \varepsilon^{-1}\|g\|_{L^2(\mathbb{R}^N)}\|\nabla_y\rho\|_{L^\infty(Y';L^2(Y)^N)}\right\}.$$

Proof. The idea is to consider the product space $L^2(Y'; L^2_\#(Y))$ and expand $\rho(y;\eta)$ in two steps. First using the fact that $\{\bar{\phi}_m(\cdot;\eta)\}_{m=1}^\infty$ is an orthonormal basis in $L^2_\#(Y)$, we get

$$\rho(y;\eta) = \sum_{m=1}^\infty a_m(\eta)\bar{\phi}_m(y;\eta) \quad \forall y \in Y,\ \eta \in Y'.$$

Next, for each m, we can expand $a_m(\eta)$ in the usual Fourier series:

$$a_m(\eta) = \sum_{n\in\mathbb{Z}^N} a_{mn} e^{2\pi i n\cdot\eta} \quad \forall \eta \in Y'.$$

The corresponding Parseval's identities are as follows:

$$\|\rho(\cdot;\eta)\|^2_{L^2(Y)} = \sum_m |a_m(\eta)|^2 \quad \forall \eta \in Y',$$
$$\int_{Y'} |a_m(\eta)|^2 d\eta = \sum_{n\in\mathbb{Z}^N} |a_{mn}|^2 \quad \forall m \in \mathbb{N}.$$

Using this expansion, we can rewrite $J^\varepsilon g$ as follows:

$$J^\varepsilon g(\xi) = \sum_{m=1}^\infty \sum_{n\in\mathbb{Z}^N} a_{mn} e^{2\pi i \varepsilon n\cdot\xi} \int_{\mathbb{R}^N} g(x) e^{-ix\cdot\xi} \bar{\phi}_m\left(\frac{x}{\varepsilon};\varepsilon\xi\right) dx,$$

which, according to the definition of $B^\varepsilon_m g(\xi)$, is equal to

$$J^\varepsilon g(\xi) = \sum_{m=1}^\infty \sum_{n\in\mathbb{Z}^N} a_{mn} e^{2\pi i \varepsilon n\cdot\xi} B^\varepsilon_m g(\xi) = \sum_{m=1}^\infty a_m(\varepsilon\xi) B^\varepsilon_m g(\xi).$$

By the Cauchy–Schwarz inequality,

$$\begin{aligned}|J^\varepsilon g(\xi)|^2 &\le \left(\sum_{m=1}^\infty |a_m(\varepsilon\xi)|^2\right)\left(\sum_{m=1}^\infty |B^\varepsilon_m g(\xi)|^2\right)\\ &= \|\rho(\cdot;\varepsilon\xi)\|^2_{L^2(Y)} \left(\sum_{m=1}^\infty |B^\varepsilon_m g(\xi)|^2\right)\\ &\le \|\rho\|^2_{L^\infty(Y';L^2_\#(Y))} \left(\sum_{m=1}^\infty |B^\varepsilon_m g(\xi)|^2\right).\end{aligned}$$

The proof of (i) is complete if we integrate the above inequality with respect to $\xi \in \varepsilon^{-1}Y'$ and apply Theorem 1.3. For the proof of (ii), we multiply (2.8) by $(-i\xi_k)$ and obtain

$$(-i\xi_k) J^\varepsilon g(\xi) = \int_{\mathbb{R}^N} g(x)(-i\xi_k) e^{-ix\cdot\xi} \rho\left(\frac{x}{\varepsilon};\varepsilon\xi\right) dx,$$

which, by integration by parts, can be rewritten as

$$(-i\xi_k) J^\varepsilon g(\xi) = -\int_{\mathbb{R}^N} \frac{\partial g}{\partial x_k}(x) e^{-ix\cdot\xi} \rho\left(\frac{x}{\varepsilon};\varepsilon\xi\right) dx - \varepsilon^{-1} \int_{\mathbb{R}^N} g(x) e^{-ix\cdot\xi} \frac{\partial \rho}{\partial y_k}\left(\frac{x}{\varepsilon};\varepsilon\xi\right) dx.$$

It is now sufficient to apply (i) to each of the terms on the right-hand side of the above relation. □

Next, we will need some properties of the classical *discrete Fourier transform* in our asymptotic description of the first Bloch transform. In particular, we are interested in the relation between discrete and continuous Fourier transforms. To this end, let us begin by introducing some necessary notations. Let $\{Y_\ell^\varepsilon\}_{\ell\in\mathbb{Z}^N}$ be the mesh of $\mathbb{R}^N$ generated by the cell εY. More precisely, $Y_\ell^\varepsilon = x_\ell^\varepsilon + \varepsilon Y$ where $x_\ell^\varepsilon = 2\pi\varepsilon\ell$ is the origin of the cell Y_ℓ^ε. We recall the definition of the discrete Fourier transform of a function corresponding to this mesh: Let $p > N$ be given. For $g \in W^{1,p}(\mathbb{R}^N)$ with compact support we define

$$F^\varepsilon g(\xi) = \sum_{\ell\in\mathbb{Z}^N} g(x_\ell^\varepsilon) e^{-ix_\ell^\varepsilon\cdot\xi} \quad \forall \xi \in \varepsilon^{-1}Y'. \tag{2.9}$$

It is worthwhile to recall that $W^{1,p}(\mathbb{R}^N)$ is embedded in $\mathcal{C}^0(\mathbb{R}^N)$ when $p > N$, and so $g(x_\ell^\varepsilon)$ is well defined.

Lemma 2.6. *For $g \in W^{1,p}(\mathbb{R}^N)$ $(p > N)$ with compact support K, we have*

(i) $\varepsilon^N(\chi_{\varepsilon^{-1}Y'} F^\varepsilon g)(\xi) \to \frac{1}{(2\pi)^{N/2}}\widehat{g}(\xi)$ *for* $\xi \in \mathbb{R}^N$.

(ii) $\|\varepsilon^N F^\varepsilon g\|_{L^2(\varepsilon^{-1}Y')} \le c|K|^{\frac{p-2}{2p}}\{\|g\|_{L^p(\mathbb{R}^N)} + \varepsilon\|\nabla g\|_{L^p(\mathbb{R}^N)^N}\}$, $|K|$ *denotes the measure of* K.

(iii) $\varepsilon^N\chi_{\varepsilon^{-1}Y'} F^\varepsilon g \to \frac{1}{(2\pi)^{N/2}}\widehat{g}$ *in* $L^2(\mathbb{R}^N)$.

Proof. To prove (i), we multiply (2.9) by ε^N to get

$$\varepsilon^N F^\varepsilon g(\xi) = \frac{1}{(2\pi)^N}\sum_{\ell\in\mathbb{Z}^N} g(x_\ell^\varepsilon)e^{-ix_\ell^\varepsilon\cdot\xi}|Y_\ell^\varepsilon|.$$

We regard the right-hand side of the above equality as a Riemann sum of the integral

$$\frac{1}{(2\pi)^N}\int_{\mathbb{R}^N} g(x)e^{-ix\cdot\xi}dx$$

which converges to it as $\varepsilon \to 0$.

To prove (ii), we observe that the right-hand side of (2.9) is nothing but the Fourier series in the variable $\xi \in \varepsilon^{-1}Y'$. Therefore, by Parseval's identity, we get

$$\varepsilon^N\int_{\varepsilon^{-1}Y'} |F^\varepsilon g(\xi)|^2 d\xi = \sum_{\ell\in\mathbb{Z}^N} |g(x_\ell^\varepsilon)|^2.$$

We multiply this relation by ε^N and rewrite it as

$$\varepsilon^{2N}\int_{\varepsilon^{-1}Y'} |F^\varepsilon g(\xi)|^2 d\xi = \frac{1}{(2\pi)^N}\sum_{\ell\in\mathbb{Z}^N} |g(x_\ell^\varepsilon)|^2|Y_\ell^\varepsilon|. \tag{2.10}$$

To estimate the right-hand side of the above equality, we integrate the inequality

$$|g(x_\ell^\varepsilon)|^2 \le 2\left\{|g(x)|^2 + |g(x) - g(x_\ell^\varepsilon)|^2\right\}, \quad x \in Y_\ell^\varepsilon,$$

over Y_ℓ^ε to obtain

$$|g(x_\ell^\varepsilon)|^2|Y_\ell^\varepsilon| \le 2\left\{\int_{Y_\ell^\varepsilon} |g(x)|^2 dx + \int_{Y_\ell^\varepsilon} |g(x) - g(x_\ell^\varepsilon)|^2 dx\right\}. \tag{2.11}$$

Since $p > N$, we can use the classical Morrey's inequality (see Brezis [7, p. 167]) to deduce

$$|g(x) - g(x^\varepsilon_\ell)| \leq c\varepsilon^{1-\frac{N}{p}} |\nabla g|_{L^p(Y^\varepsilon_\ell)^N}.$$

Using both the above estimate in (2.11) and the Hölder inequality and summing over $\ell \in \mathbb{Z}^N$, we complete the proof of (ii).

To prove the statement (iii) we first use (i) and (ii) to deduce that

$$\varepsilon^N \chi_{\varepsilon^{-1}Y'}(\xi) F^\varepsilon g(\xi) \rightharpoonup \frac{1}{(2\pi)^{N/2}} \widehat{g}(\xi) \quad \text{in } L^2(\mathbb{R}^N)\text{-weak.}$$

Let us now expand

$$\begin{aligned} \left\| \varepsilon^N \chi_{\varepsilon^{-1}Y'} F^\varepsilon g - \tfrac{1}{(2\pi)^{N/2}} \widehat{g} \right\|^2_{L^2(\mathbb{R}^N)} &= \varepsilon^{2N} \|F^\varepsilon g\|^2_{L^2(\mathbb{R}^N)} - \tfrac{2\varepsilon^N}{(2\pi)^{N/2}} (\chi_{\varepsilon^{-1}Y'} F^\varepsilon g, \widehat{g}) \\ &\quad + \tfrac{1}{(2\pi)^N} \|\widehat{g}\|^2_{L^2(\mathbb{R}^N)}. \end{aligned}$$

Now relation (2.10) shows that

$$\varepsilon^{2N} \|F^\varepsilon g\|^2_{L^2(\mathbb{R}^N)} \to \frac{1}{(2\pi)^N} \int_{\mathbb{R}^N} |g|^2 dx = \frac{1}{(2\pi)^N} \|\widehat{g}\|^2_{L^2(\mathbb{R}^N)}.$$

Thanks to the above weak convergence, the second term converges to

$$-\frac{2}{(2\pi)^{N/2}} \frac{1}{(2\pi)^{N/2}} \|\widehat{g}\|^2_{L^2(\mathbb{R}^N)}.$$

This simple computation establishes the strong convergence in $L^2(\mathbb{R}^N)$. □

3. Higher Bloch modes are negligible. In this section, we consider a sequence of solutions u^ε of the equation with $f \in H^{-1}(\mathbb{R}^N)$:

$$A^\varepsilon u^\varepsilon = f \quad \text{in} \quad \mathbb{R}^N, \quad u^\varepsilon \in H^1(\mathbb{R}^N). \tag{3.1}$$

Let us recall that the above equation is equivalent to (1.20) in the Bloch space. In what follows, we present a systematic method of obtaining estimates on the solution in Sobolev spaces L^2 and H^1. In particular, we show that the component of u^ε in the higher Bloch modes does not play any role in the analysis of correctors of first and second order provided f is sufficiently smooth. Thus we consider v^ε defined in (1.21), which is nothing but the projection of u^ε corresponding to all higher Bloch modes. Estimates on v^ε derived in this section improve Proposition 1.4.

PROPOSITION 3.1. *We have the following estimates for $f \in L^2(\mathbb{R}^N)$:*

(i) $|v^\varepsilon|_{H^1(\mathbb{R}^N)} \leq c\varepsilon \|f\|_{L^2(\mathbb{R}^N)}$,

(ii) $\|v^\varepsilon\|_{L^2(\mathbb{R}^N)} \leq c\varepsilon \|f\|_{H^{-1}(\mathbb{R}^N)}$.

Proof. To show (i), we apply Lemma 2.1 with $g = v^\varepsilon$ and use (1.20). We obtain

$$\begin{aligned} \|\nabla v^\varepsilon\|^2_{L^2(\mathbb{R}^N)^N} &\leq c \int_{\varepsilon^{-1}Y'} \sum_{m=2}^{\infty} \frac{1}{\lambda^\varepsilon_m(\xi)} |B^\varepsilon_m f(\xi)|^2 d\xi \\ &\leq c \sup_{m \geq 2,\ \xi \in \varepsilon^{-1}Y'} \frac{1}{\lambda^\varepsilon_m(\xi)} \|f\|^2_{L^2(\mathbb{R}^N)}. \end{aligned}$$

Proof of (i) is complete since we have (cf. (1.25))

$$\sup_{m\geq 2,\ \xi\in\varepsilon^{-1}Y'} \frac{1}{\lambda_m^\varepsilon(\xi)} \leq \frac{1}{\lambda_2^{(N)}}\varepsilon^2. \tag{3.2}$$

For the proof of (ii), we apply Lemma 2.2 with $g = f$ and (1.20). We have

$$\begin{aligned}\|v^\varepsilon\|^2_{L^2(\mathbb{R}^N)} &= \int_{\varepsilon^{-1}Y'} \sum_{m=2}^\infty |B_m^\varepsilon u^\varepsilon(\xi)|^2 d\xi \\ &= \int_{\varepsilon^{-1}Y'} \sum_{m=2}^\infty \frac{1}{\lambda_m^\varepsilon(\xi)^2} |B_m^\varepsilon f(\xi)|^2 d\xi.\end{aligned}$$

Writing

$$\frac{1}{\lambda_m^\varepsilon(\xi)^2}|B_m^\varepsilon f(\xi)|^2 = \frac{1+\lambda_m^\varepsilon(\xi)}{\lambda_m^\varepsilon(\xi)^2}\frac{|B_m^\varepsilon f(\xi)|^2}{1+\lambda_m^\varepsilon(\xi)}$$

and using (3.2), we deduce that

$$\frac{1}{\lambda_m^\varepsilon(\xi)^2}|B_m^\varepsilon f(\xi)|^2 \leq c\varepsilon^2 \frac{|B_m^\varepsilon f(\xi)|^2}{1+\lambda_m^\varepsilon(\xi)}.$$

The proof is complete if we use Lemma 2.2. □

While the above proposition shows that v^ε can be neglected at the level of the first order correctors (cf. (1.10)), the next result will demonstrate that v^ε can be neglected at the level of correctors of first and second order. These finer estimates require naturally higher order regularity of f but not of the coefficients $a_{k\ell}(y)$. Let us state the following proposition, whose proof is similar to the previous one and hence will not be repeated.

Proposition 3.2. *We have the following estimates for $f \in H^1(\mathbb{R}^N)$:*

(i) $|v^\varepsilon|_{H^1(\mathbb{R}^N)} \leq c\varepsilon^2\|f\|_{H^1(\mathbb{R}^N)}$,

(ii) $\|v^\varepsilon\|_{L^2(\mathbb{R}^N)} \leq c\varepsilon^2\|f\|_{L^2(\mathbb{R}^N)}$. □

Assuming $a_{k\ell}$ are in $W^{1,\infty}_\#(Y)$ and further assumptions, we can obtain H^2-estimates on the solution. This is difficult as it involves more subtleties (see [12]).

4. Taylor expansion of the first Bloch eigenvalue and eigenvector. The purpose of this section is to indicate a systematic method to compute derivatives of the first Bloch eigenvalue $\lambda_1(\eta)$ and the first Bloch eigenvector $\phi_1(\cdot;\eta)$ at $\eta = 0$. In particular, we will prove Propositions 1.9 and 1.10. Recall that $\lambda_1(\eta)$ and $\phi_1(\cdot;\eta)$ depend analytically on η in a small neighborhood B_δ of $\eta = 0$. At the cost of reducing this neighborhood, we claim that the branch $\eta \mapsto \phi_1(\cdot;\eta)$ can be chosen so that the following conditions are satisfied simultaneously:

$$\eta \in B_\delta \mapsto \phi_1(\cdot;\eta) \in H^1_\#(Y) \quad \text{is analytic,} \tag{4.1}$$

$$\|\phi_1(\cdot;\eta)\|_{L^2(Y)} = 1 \quad \forall \eta \in B_\delta, \tag{4.2}$$

$$\mathfrak{Im}\int_Y \phi_1(y;\eta)dy = 0 \quad \forall \eta \in B_\delta. \tag{4.3}$$

In what follows, we will see that the above conditions uniquely fix the eigenvector $\phi_1(\cdot;\eta)$. We remark that the condition (4.2) is classical, whereas the condition (4.3) is somewhat unusual and can be achieved as indicated below. The idea consists of multiplying $\phi_1(\cdot;\eta)$ by a complex number $(\alpha_1(\eta)+i\alpha_2(\eta))$ where $\alpha_1(\eta)$ and $\alpha_2(\eta)$ are real analytic with respect to η and are chosen such that

$$\mathfrak{Im}\int_Y (\alpha_1(\eta)+i\alpha_2(\eta))\phi_1(y;\eta)dy = 0.$$

If we define

$$d(\eta) = (d_1(\eta), d_2(\eta)) \overset{\text{def}}{=} \left(\mathfrak{Im}\int_Y \phi_1(y;\eta)dy, \mathfrak{Re}\int_Y \phi_1(y;\eta)dy\right),$$

then the above condition is equivalent to

$$\alpha_1(\eta)d_1(\eta) + \alpha_2(\eta)d_2(\eta) = 0 \quad \forall \eta \in B_\delta.$$

Obviously, one such choice which is analytic is as follows:

$$\alpha_1(\eta) = -d_2(\eta), \quad \alpha_2(\eta) = d_1(\eta).$$

Of course, the above procedure has destroyed condition (4.2) (but not condition (4.1)). However, it can be regained by dividing by $|d(\eta)|$. This is possible because $d(0) \neq 0$ by our choice of $\phi_1(\cdot;0)$ (see Proposition 1.5).

Thanks to our choice of the branch satisfying (4.1)–(4.3), we will now draw some consequences which will simplify the computations below. In fact, differentiating (4.2) with respect to η, we successively get for all $k,\ell,m,n = 1,\ldots,N$

$$\mathfrak{Re}\langle D_k\phi_1(\cdot;\eta), \phi_1(\cdot;\eta)\rangle = 0, \tag{4.4}$$

$$\mathfrak{Re}\langle D^2_{k\ell}\phi_1(\cdot;\eta), \phi_1(\cdot;\eta)\rangle + \mathfrak{Re}\langle D_k\phi_1(\cdot;\eta), D_\ell\phi_1(\cdot;\eta)\rangle = 0, \tag{4.5}$$

$$\left\{\begin{aligned} &\mathfrak{Re}\langle D^3_{k\ell m}\phi_1(\cdot;\eta), \phi_1(\cdot;\eta)\rangle + \mathfrak{Re}\langle D^2_{k\ell}\phi_1(\cdot;\eta), D_m\phi_1(\cdot;\eta)\rangle \\ &\quad + \mathfrak{Re}\langle D^2_{km}\phi_1(\cdot;\eta), D_\ell\phi_1(\cdot;\eta)\rangle + \mathfrak{Re}\langle D_k\phi_1(\cdot;\eta), D^2_{\ell m}\phi_1(\cdot;\eta)\rangle = 0,\end{aligned}\right. \tag{4.6}$$

$$\left\{\begin{aligned} &\mathfrak{Re}\langle D^4_{k\ell mn}\phi_1(\cdot;\eta), \phi_1(\cdot;\eta)\rangle + \mathfrak{Re}\langle D^3_{k\ell m}\phi_1(\cdot;\eta), D_n\phi_1(\cdot;\eta)\rangle \\ &\quad + \mathfrak{Re}\langle D^3_{k\ell n}\phi_1(\cdot;\eta), D_m\phi_1(\cdot;\eta)\rangle + \mathfrak{Re}\langle D^2_{k\ell}\phi_1(\cdot;\eta), D^2_{mn}\phi_1(\cdot;\eta)\rangle \\ &\quad\quad + \mathfrak{Re}\langle D^3_{kmn}\phi_1(\cdot;\eta), D_\ell\phi_1(\cdot;\eta)\rangle + \mathfrak{Re}\langle D^2_{km}\phi_1(\cdot;\eta), D^2_{\ell n}\phi_1(\cdot;\eta)\rangle \\ &\quad\quad\quad + \mathfrak{Re}\langle D^2_{kn}\phi_1(\cdot;\eta), D^2_{\ell m}\phi_1(\cdot;\eta)\rangle + \mathfrak{Re}\langle D_k\phi_1(\cdot;\eta), D^3_{\ell mn}\phi_1(\cdot;\eta)\rangle = 0,\end{aligned}\right. \tag{4.7}$$

where $\langle\cdot;\cdot\rangle$ denotes the scalar product in $L^2_\#(Y)$. On the other hand, differentiation of (4.3) yields

$$\mathfrak{Im}\int_Y D^\beta\phi_1(y;\eta)dy = 0 \quad \forall\ \beta \in \mathbb{Z}^N_+. \tag{4.8}$$

From these sets of relations, it follows that

$$\int_Y D^\beta\phi_1(y;0)dy = 0 \quad \forall\ \beta \in \mathbb{Z}^N_+ \text{ with } |\beta| \text{ odd.} \tag{4.9}$$

4.1. First order derivatives. If we differentiate the eigenvalue equation $(A(\eta) - \lambda_1(\eta))\phi_1(\cdot;\eta) = 0$ once with respect to η_k, we obtain

$$D_k(A(\eta) - \lambda_1(\eta))\phi_1(\cdot;\eta) + (A(\eta) - \lambda_1(\eta))D_k\phi_1(\cdot;\eta) = 0. \tag{4.10}$$

Taking the scalar product with $\phi_1(\cdot;\eta)$, we get

$$\langle [D_k(A - \lambda_1)]\phi_1, \phi_1\rangle = 0, \tag{4.11}$$

where we have suppressed the dependence on η for ease of writing. We will continue with this convention in what follows provided there is no ambiguity. It follows from (1.18) that

$$D_kA(0) = iC_k \quad \forall \eta \in Y', \tag{4.12}$$

where the operator C_k is defined in (1.19).

If we evaluate the relation (4.11) at $\eta = 0$ and use the structure of C_k, we immediately get that

$$D_k\lambda_1(0) = 0 \quad \forall k = 1, \ldots, N. \tag{4.13}$$

The next step is to compute the first order derivatives of ϕ_1 at $\eta = 0$. To this end, we go back to (4.10) and use (4.13). We obtain

$$AD_k\phi_1(\cdot;0) = -D_kA(0)\phi_1(\cdot;0) = -iC_k\phi_1(\cdot;0).$$

Taking into account (4.9) and the above equation, we can solve for $D_k\phi_1(y;0)$ and obtain

$$D_k\phi_1(y;0) = i\phi_1(y;0)\chi_k(y) = ip^{(0)}\chi_k(y), \tag{4.14}$$

where, we recall, χ_k satisfies (1.13) and the constant $p^{(0)}$ was fixed in Proposition 1.5. Thus, the first order derivative is completely determined and

$$D_k\phi_1(y;0) \quad \text{is purely imaginary.} \tag{4.15}$$

4.2. Second order derivatives. Our starting point is the relation (4.10), which we differentiate once with respect to η. We obtain

$$[D^2_{k\ell}(A - \lambda_1)]\phi_1 + [D_k(A - \lambda_1)]D_\ell\phi_1 + [D_\ell(A - \lambda_1)]D_k\phi_1 + (A - \lambda_1)D^2_{k\ell}\phi_1 = 0. \tag{4.16}$$

Taking the scalar product with ϕ_1, we get

$$\langle [D^2_{k\ell}(A - \lambda_1)]\phi_1, \phi_1\rangle + \langle [D_k(A - \lambda_1)]D_\ell\phi_1, \phi_1\rangle + \langle [D_\ell(A - \lambda_1)]D_k\phi_1, \phi_1\rangle = 0 \tag{4.17}$$

for all $\eta \in B_\delta$. If we use the information obtained in section 4.1 on $D_k\lambda_1(0)$, $D_k\phi_1(\cdot;0)$, $D_kA(0)$, and

$$D^2_{k\ell}A(\eta) = 2a_{k\ell}(y) \quad \forall k, \ell = 1, \ldots, N, \ \eta \in Y', \tag{4.18}$$

we obtain

$$\frac{1}{2!}D^2_{k\ell}\lambda_1(0) = \frac{1}{|Y|}\int_Y a_{k\ell}(y)dy - \frac{1}{2|Y|}\int_Y (C_k\chi_\ell(y) + C_\ell\chi_k(y))dy$$

$$= \frac{1}{2}(q_{k\ell} + q_{\ell k}) = q_{k\ell} \quad \forall k,\ell = 1,\dots,N. \tag{4.19}$$

As before, the next step is to compute $D^2_{k\ell}\phi_1(\cdot;0)$. For this purpose, we go back to (4.16) and rewrite it with $\eta = 0$ as follows:

$$AD^2_{k\ell}\phi_1(\cdot;0) = \left\{-2(a_{k\ell} - q_{k\ell}) + C_k\chi_\ell + C_\ell\chi_k\right\}\phi_1(\cdot;0).$$

By comparing the above equation with (1.30) and using the simplicity of the eigenvalue under consideration, we see that $D^2_{k\ell}\phi_1(\cdot;0)$ is of the form

$$\frac{1}{2!}D^2_{k\ell}\phi_1(y;0) = -p^{(0)}\chi_{k\ell}(y) - \beta^{(2)}_{k\ell}p^{(0)}$$

for some constant $\beta^{(2)}_{k\ell}$. Thanks to (4.5) and (4.8), we can infer that

$$\beta^{(2)}_{k\ell} \quad \text{and} \quad D^2_{k\ell}\phi_1(\cdot;0) \quad \text{are real.} \tag{4.20}$$

Moreover, $\beta^{(2)}_{k\ell}$ admits the expression given in Proposition 1.10.

4.3. Third order derivatives. From the calculations done so far, it is now clear how to proceed further to calculate higher order derivatives. Therefore we will be brief here. Differentiating (4.16), we get

$$\begin{cases}[D^3_{k\ell m}(A-\lambda_1)]\phi_1 + [D^2_{k\ell}(A-\lambda_1)]D_m\phi_1 + [D^2_{\ell m}(A-\lambda_1)]D_k\phi_1 \\ \quad + [D^2_{km}(A-\lambda_1)]D_\ell\phi_1 + [D_k(A-\lambda_1)]D^2_{\ell m}\phi_1 + [D_\ell(A-\lambda_1)]D^2_{km}\phi_1 \\ \quad\quad + [D_m(A-\lambda_1)]D^2_{k\ell}\phi_1 + (A-\lambda_1)D^3_{k\ell m}\phi_1 = 0.\end{cases} \tag{4.21}$$

Taking the scalar product with ϕ_1, we get

$$\begin{cases}\langle[D^3_{k\ell m}(A-\lambda_1)]\phi_1,\phi_1\rangle + \langle[D^2_{k\ell}(A-\lambda_1)]D_m\phi_1,\phi_1\rangle + \langle[D^2_{\ell m}(A-\lambda_1)]D_k\phi_1,\phi_1\rangle \\ \quad + \langle[D^2_{km}(A-\lambda_1)]D_\ell\phi_1,\phi_1\rangle + \langle[D_k(A-\lambda_1)]D^2_{\ell m}\phi_1,\phi_1\rangle \\ \quad\quad + \langle[D_\ell(A-\lambda_1)]D^2_{km}\phi_1,\phi_1\rangle + \langle[D_m(A-\lambda_1)]D^2_{k\ell}\phi_1,\phi_1\rangle = 0.\end{cases} \tag{4.22}$$

To conclude that $D^3_{k\ell m}\lambda_1(0) = 0$, it is enough to use the following information in the above relation:

$$\begin{cases} D_kA \text{ is purely imaginary}, \quad D^2_{k\ell}A \text{ is real}, \quad D^3_{k\ell m}A = 0, \\ \phi_1(0), D^2_{k\ell}\phi_1(0) \text{ are real}, \quad D_k\phi_1(0) \text{ is purely imaginary.}\end{cases} \tag{4.23}$$

It is evident that the above argument is very general and so can be used to establish that all odd order derivatives of λ_1 at $\eta = 0$ vanish. This proves the first part of Proposition 1.9.

To find the third order derivatives of ϕ_1 at $\eta = 0$, we realize that (4.21) defines a periodic problem for $D^3_{k\ell m}\phi_1(\cdot;0)$ which can be compared with (1.31). Further, the relation (4.9) says that its average vanishes. These observations are enough to get the expression of $D^3_{k\ell m}\phi_1(\cdot;0)$ given in Proposition 1.10. We conclude by observing the following important property:

$$D^3_{k\ell m}\phi_1(y;0) \quad \text{is purely imaginary.} \tag{4.24}$$

4.4. Fourth order derivatives. To arrive at the expressions for the fourth order derivatives of λ_1 and ϕ_1 at $\eta = 0$ given in Propositions 1.9 and 1.10, we follow the same arguments as in section 4.3.

5. Convergence of the first Bloch transform to the Fourier transform.

This section is devoted to the proof of the next proposition which shows the sense in which the Fourier transform is approximated by the first Bloch transform.

PROPOSITION 5.1.

(i) *For every $g \in L^2(\mathbb{R}^N)$ with compact support, we have*

$$\chi_{\varepsilon^{-1}Y'}(\xi)B_1^\varepsilon g(\xi) \to \widehat{g}(\xi) \quad \textit{in} \quad L^\infty_{loc}(\mathbb{R}^N_\xi).$$

(ii) *If $g \in L^2(\mathbb{R}^N)$, we have*

$$\chi_{\varepsilon^{-1}Y'}(\xi)B_1^\varepsilon g(\xi) \to \widehat{g}(\xi) \quad \textit{in} \quad L^2(\mathbb{R}^N_\xi).$$

This will be a consequence of a more general result. In order to state it, we need to introduce some new notations. We associate with every function $\rho = \rho(y;\eta)$ defined on $Y \times Y'$ which is Y-periodic in y the following function:

$$\widetilde{\rho}^{(0)}(\eta) = \frac{1}{|Y|}\int_Y \rho(y;\eta)e^{-iy\cdot\eta}dy, \quad \eta \in Y'. \tag{5.1}$$

With this notation, we have the following proposition.

PROPOSITION 5.2. *We suppose $\rho \in L^\infty(Y';L^2_\#(Y))$. Then for all $g \in W^{1,p}(\mathbb{R}^N)$ with compact support K and with $p > N$, we have*

$$\chi_{\varepsilon^{-1}Y'}(\xi)\left(J^\varepsilon g(\xi) - (2\pi)^{N/2}\tilde{\rho}^{(0)}(\varepsilon\xi)\widehat{g}(\xi)\right) \to 0 \quad \textit{in} \quad L^2(\mathbb{R}^N_\xi), \tag{5.2}$$

where, we recall, $J^\varepsilon g$ was defined in (2.8).

The proof will be taken up later. Admitting it for the moment, we turn our attention to the following proof.

Proof of Proposition 5.1. If $g \in L^2(\mathbb{R}^N)$ with compact support K, we have for all $\xi \in \mathbb{R}^N$

$$\begin{aligned}|\chi_{\varepsilon^{-1}Y'}(\xi)B_1^\varepsilon g(\xi) - \widehat{g}(\xi)| &\le |\chi_{\varepsilon^{-1}Y'}(\xi)(B_1^\varepsilon g(\xi) - \widehat{g}(\xi))| + |(\chi_{\varepsilon^{-1}Y'}(\xi) - 1)\widehat{g}(\xi)| \\ &\le c|K|\|g\|_{L^2(\mathbb{R}^N)}\|\phi_1(\cdot;\varepsilon\xi) - \phi_1(\cdot;0)\|_{L^2(Y)} + |(\chi_{\varepsilon^{-1}Y'}(\xi) - 1)\widehat{g}(\xi)|.\end{aligned}$$

If $|\xi|$ is bounded, then by using the fact that the map $\eta \mapsto \phi_1(\cdot;\eta) \in L^2_\#(Y)$ is Lipschitz near $\eta = 0$, we deduce

$$||\phi_1(\cdot;\varepsilon\xi) - \phi_1(\cdot;0)\|_{L^2(Y)} \le c\varepsilon.$$

This completes the proof of (i).

The proof of (ii) is more involved. First, according to Theorem 1.3, we have the uniform estimate

$$\int_{\varepsilon^{-1}Y'} |B_1^\varepsilon g(\xi)|^2 d\xi \leq \int_{\mathbb{R}^N} |g(x)|^2 dx,$$

and so, by the usual density arguments, it is enough to prove (ii) with $g \in \mathcal{D}(\mathbb{R}^N)$. We can now complete the proof using Proposition 5.2. Indeed, with $\rho = \bar{\phi}_1$, we see that

$$\tilde{\rho}^{(0)}(\varepsilon\xi) \to p^{(0)} \quad \text{and} \quad B_1^\varepsilon g(\xi) = J^\varepsilon g(\xi) \quad \forall \xi \in \mathbb{R}^N,$$

which implies, by Lebesgue's dominated convergence theorem, that

$$(2\pi)^{N/2} \chi_{\varepsilon^{-1}Y'}(\xi) \tilde{\rho}^{(0)}(\varepsilon\xi) \widehat{g}(\xi) \to \widehat{g}(\xi) \quad \text{in} \quad L^2(\mathbb{R}^N_\xi). \qquad \square$$

Proof of Proposition 5.2. The key point is that the variation of $\rho(\frac{x}{\varepsilon}; \varepsilon\xi)$ with respect to x is faster than that of g. To exploit this, we consider the ε-mesh $\{Y_\ell^\varepsilon\}_{\ell \in \mathbb{Z}^N}$ generated by the cell εY which was already introduced at the end of section 2. We decompose

$$\begin{aligned} J^\varepsilon g(\xi) &= \sum_{\ell \in \mathbb{Z}^N} \int_{Y_\ell^\varepsilon} g(x) e^{-ix\cdot\xi} \rho\left(\frac{x}{\varepsilon}; \varepsilon\xi\right) dx \\ &= \sum_{\ell \in \mathbb{Z}^N} g(x_\ell^\varepsilon) \int_{Y_\ell^\varepsilon} e^{-ix\cdot\xi} \rho\left(\frac{x}{\varepsilon}; \varepsilon\xi\right) dx + r_1^\varepsilon(\xi), \end{aligned} \tag{5.3}$$

where

$$r_1^\varepsilon(\xi) = \sum_{\ell \in \mathbb{Z}^N} \int_{Y_\ell^\varepsilon} (g(x) - g(x_\ell^\varepsilon)) e^{-ix\cdot\xi} \rho\left(\frac{x}{\varepsilon}; \varepsilon\xi\right) dx. \tag{5.4}$$

The first term on the right-hand side of (5.3) can be, by means of the change of variables $x = x_\ell^\varepsilon + \varepsilon y$, transformed into

$$|Y| \varepsilon^N F^\varepsilon g(\xi) \widetilde{\rho}^{(0)}(\varepsilon\xi),$$

where $F^\varepsilon g$ is the discrete Fourier transform of g and $\tilde{\rho}^{(0)}$ is defined in (5.1). Since we know that $\chi_{\varepsilon^{-1}Y'}(\xi) \varepsilon^N F^\varepsilon g(\xi) \to \frac{1}{(2\pi)^{N/2}} \hat{g}(\xi)$ in $L^2(\mathbb{R}^N)$ (cf. Lemma 2.6), our hypothesis on ρ ensures that

$$\left\| \chi_{\varepsilon^{-1}Y'}(\xi) \left\{ |Y| \varepsilon^N F^\varepsilon g(\xi) - (2\pi)^{N/2} \widehat{g}(\xi) \right\} \tilde{\rho}^{(0)}(\varepsilon\xi) \right\|_{L^2(\mathbb{R}^N)} \to 0. \tag{5.5}$$

Thus, to complete the proof, it is enough to show that

$$\|r_1^\varepsilon\|_{L^2(\varepsilon^{-1}Y')} \leq \frac{c(K)}{(1 - \frac{N}{p})} \varepsilon \|\rho\|_{L^\infty(Y'; L^2_\#(Y))} \|\nabla g\|_{L^p(\mathbb{R}^N)}. \tag{5.6}$$

To this end, we rewrite r_1^ε in a slightly different form, namely,

$$r_1^\varepsilon(\xi) = \int_{\mathbb{R}^N} \widetilde{g}_1^\varepsilon(x) e^{-ix\cdot\xi} \rho\left(\frac{x}{\varepsilon}; \varepsilon\xi\right) dx, \tag{5.7}$$

where

$$\widetilde{g}_1^\varepsilon(x) = \sum_{\ell\in\mathbb{Z}^N} (g(x) - g(x_\ell^\varepsilon))\chi_{Y_\ell^\varepsilon}(x). \tag{5.8}$$

We already know how to estimate integrals of the type (5.7) in $L^2(\mathbb{R}^N)$ (see Lemma 2.5), and so we can deduce (5.6) provided we have the estimate

$$\|\widetilde{g}_1^\varepsilon\|_{L^2(\mathbb{R}^N)} \le \frac{c(K)}{(1-\frac{N}{p})}\varepsilon\|\nabla g\|_{L^p(\mathbb{R}^N)}. \tag{5.9}$$

Thanks to our hypothesis, we can deduce a stronger estimate, namely,

$$\|\widetilde{g}_1^\varepsilon\|_{L^p(\mathbb{R}^N)} \le \frac{c}{(1-\frac{N}{p})}\varepsilon\|\nabla g\|_{L^p(\mathbb{R}^N)}, \tag{5.10}$$

where c is a constant independent of K, the support of g. We note that (5.10) is a simple consequence of Morrey's estimate (see [7, p. 167]).

Finally, we note that (5.9) can be obtained from (5.10) with $c(K) = c|K|^{1-\frac{2}{p}}$ and a simple application of the Hölder inequality. □

The proof of Proposition 5.2 shows that the result can be strengthened by assuming suitable smoothness on g. Our next result is an example in this direction. It introduces naturally the following quantities:

$$\widetilde{\rho}^{(k)}(\eta) = \frac{1}{|Y|}\int_Y \rho(y;\eta)y_k e^{-iy\cdot\eta}dy \quad \forall k = 1,\dots,N, \quad \eta\in Y'. \tag{5.11}$$

Then we have the following corrector result for $J^\varepsilon g$.

PROPOSITION 5.3. *We suppose $\rho\in L^\infty(Y';L^2_\#(Y))$. Then for all $g\in W^{2,p}(\mathbb{R}^N)$ with compact support K and with $p>N$, we have*

$$\chi_{\varepsilon^{-1}Y'}(\xi)\left\{J^\varepsilon g(\xi) - (2\pi)^{N/2}\big[\tilde{\rho}^{(0)}(\varepsilon\xi) + i\varepsilon\xi_k\tilde{\rho}^{(k)}(\varepsilon\xi)\big]\widehat{g}(\xi)\right\} \to 0 \quad \text{in} \quad L^2(\mathbb{R}^N_\xi).$$

Proof. We follow the idea of the proof of Proposition 5.2. We decompose $J^\varepsilon g(\xi)$ as

$$J^\varepsilon g(\xi) = \sum_{\ell\in\mathbb{Z}^N}\int_{Y_\ell^\varepsilon}\{g(x_\ell^\varepsilon) + \nabla g(x_\ell^\varepsilon)\cdot(x - x_\ell^\varepsilon)\}\,e^{-ix\cdot\xi}\rho\left(\frac{x}{\varepsilon};\varepsilon\xi\right)dx + r_2^\varepsilon(\xi), \tag{5.12}$$

where

$$r_2^\varepsilon(\xi) = \sum_{\ell\in\mathbb{Z}^N}\int_{Y_\ell^\varepsilon}\{g(x) - g(x_\ell^\varepsilon) - \nabla g(x_\ell^\varepsilon)\cdot(x - x_\ell^\varepsilon)\}\,e^{-ix\cdot\xi}\rho\left(\frac{x}{\varepsilon};\varepsilon\xi\right)dx. \tag{5.13}$$

We can estimate $r_2^\varepsilon(\xi)$ as follows:

$$\|r_2^\varepsilon\|_{L^2(\varepsilon^{-1}Y')} \le \frac{c(K)}{(2-\frac{N}{p})}\varepsilon^2\|\rho\|_{L^\infty(Y';L^2_\#(Y))}\,|g|_{W^{2,p}(\mathbb{R}^N)}. \tag{5.14}$$

This, in fact, will be a consequence of Lemma 2.5, because we can represent r_2^ε as follows:

$$r_2^\varepsilon(\xi) = \int_{\mathbb{R}^N}\widetilde{g}_2^\varepsilon(x)e^{-ix\cdot\xi}\rho\left(\frac{x}{\varepsilon};\varepsilon\xi\right)dx \tag{5.15}$$

with

$$\widetilde{g}_2^\varepsilon(x) = \sum_{\ell\in\mathbb{Z}^N} \left(g(x) - g(x_\ell^\varepsilon) - \nabla g(x_\ell^\varepsilon)\cdot(x - x_\ell^\varepsilon)\right)\chi_{Y_\ell^\varepsilon}(x), \tag{5.16}$$

which admits the following estimates:

$$\|\widetilde{g}_2^\varepsilon\|_{L^2(\mathbb{R}^N)} \le \frac{c(K)}{(2-\frac{N}{p})}\varepsilon^2|g|_{W^{2,p}(\mathbb{R}^N)}, \tag{5.17}$$

$$\|\widetilde{g}_2^\varepsilon\|_{L^p(\mathbb{R}^N)} \le \frac{c}{(2-\frac{N}{p})}\varepsilon^2|g|_{W^{2,p}(\mathbb{R}^N)}. \tag{5.18}$$

As before, (5.17) will be a consequence of (5.18) with $c(K) = c|K|^{1-\frac{2}{p}}$.

To establish (5.18), what we need is a generalization of Morrey's inequality for $W^{2,p}$ functions, namely,

$$|g(x)-g(x_\ell^\varepsilon)-\nabla g(x_\ell^\varepsilon)\cdot(x-x_\ell^\varepsilon)| \le \frac{c}{(2-\frac{N}{p})}|x-x_\ell^\varepsilon|^{2-\frac{N}{p}}|g|_{W^{2,p}(Y_\ell^\varepsilon)} \ \forall x\in Y_\ell^\varepsilon. \tag{5.19}$$

Admitting the above estimate, it is an easy matter to prove (5.18). But the above estimate is a consequence of Morrey's inequality for the gradient $\nabla g \in W^{1,p}(\mathbb{R}^N)$ and the following representation: for all $x \in Y_\ell^\varepsilon$,

$$g(x) - g(x_\ell^\varepsilon) - \nabla g(x_\ell^\varepsilon)\cdot(x - x_\ell^\varepsilon) = \int_0^1 \{\nabla g((1-t)x_\ell^\varepsilon + tx) - \nabla g(x_\ell^\varepsilon)\}\cdot(x - x_\ell^\varepsilon)dt.$$

This completes the proof of the estimate (5.14) on r_2^ε. Thus, as expected, r_2^ε tends to zero more rapidly. The same cannot be said for the first term on the right-hand side of (5.12). Indeed, it is equal to

$$|Y|\left[\varepsilon^N(F^\varepsilon g)(\xi)\tilde{\rho}^{(0)}(\varepsilon\xi) + \varepsilon^{N+1}\left(F^\varepsilon\frac{\partial g}{\partial x_k}\right)(\xi)\tilde{\rho}^{(k)}(\varepsilon\xi)\right]. \tag{5.20}$$

According to Lemma 2.6, we have the following convergence (apart from (5.5)):

$$\chi_{\varepsilon^{-1}Y'}\left\{|Y|\left[\varepsilon^N\left(F^\varepsilon\frac{\partial g}{\partial x_k}\right)(\xi) - \frac{1}{(2\pi)^{N/2}}i\xi_k\widehat{g}(\xi)\right]\tilde{\rho}^{(k)}(\varepsilon\xi)\right\} \to 0 \text{ in } L^2(\mathbb{R}_\xi^N). \tag{5.21}$$

This clearly allows us to complete the proof. □

6. Proof of the main convergence results. Applying the previously developed techniques and results, we are now in a position to prove the main convergence results stated in section 1.2 of the introduction (namely, Theorems 1.8 and 1.11 and the statement (1.8)). We begin by recalling briefly the set-up. We take $f \in L^2(\mathbb{R}^N)$ and consider a sequence u^ε satisfying (1.9), i.e.,

$$\begin{cases} A^\varepsilon u^\varepsilon = f & \text{in} \quad \mathbb{R}^N, \\ u^\varepsilon \rightharpoonup u^* & \text{in} \quad H^1(\mathbb{R}^N)\text{-weak}, \\ u^\varepsilon \to u^* & \text{in} \quad L^2(\mathbb{R}^N)\text{-strong}. \end{cases} \tag{6.1}$$

6.1. No concentration of energy at infinity. Our hypothesis that $u^\varepsilon \to u^*$ in $L^2(\mathbb{R}^N)$-strong may, at first sight, look artificial. But this in not the case. If Ω is bounded and smooth, then it is classical that the weak convergence in $H^1(\Omega)$ will automatically imply the strong convergence in $L^2(\Omega)$. This is not the case in $\mathbb{R}^N$. To make comparisons, the correct operator to consider is $(A^\varepsilon + I)$ instead of A^ε in $\mathbb{R}^N$. In that case, we have the following.

PROPOSITION 6.1. *Assume that w^ε satisfies*

$$\left\{ \begin{array}{rll} (A^\varepsilon + I)w^\varepsilon = g & \textit{in} & \mathbb{R}^N, \\ w^\varepsilon \rightharpoonup w^* & \textit{in} & H^1(\mathbb{R}^N)\textit{-weak}, \end{array} \right. \tag{6.2}$$

where g is a given function in $L^2(\mathbb{R}^N)$. Then

$$w^\varepsilon \to w^* \quad \textit{in} \quad L^2(\mathbb{R}^N)\textit{-strong}.$$

Proof. First of all, following the idea of Proposition 3.1, we can neglect higher Bloch modes of w^ε and w^*. More precisely, we can show

$$\int_{\varepsilon^{-1}Y'} \sum_{m=2}^{\infty} |B_m^\varepsilon w^\varepsilon(\xi)|^2 d\xi \le c\varepsilon^4,$$

$$\int_{\varepsilon^{-1}Y'} \sum_{m=2}^{\infty} |B_m^\varepsilon w^*(\xi)|^2 d\xi \le c\varepsilon^2.$$

Therefore, it remains to prove

$$\int_{\varepsilon^{-1}Y'} |B_1^\varepsilon w^\varepsilon(\xi) - B_1^\varepsilon w^*(\xi)|^2 d\xi \to 0. \tag{6.3}$$

Equation (6.2) gives the relation

$$(1 + \lambda_1^\varepsilon(\xi))B_1^\varepsilon w^\varepsilon(\xi) = B_1^\varepsilon g(\xi), \quad \xi \in \varepsilon^{-1}Y'.$$

We use it to write

$$\begin{aligned} \chi_{\varepsilon^{-1}Y'}(\xi)(B_1^\varepsilon w^\varepsilon(\xi) - B_1^\varepsilon w^*(\xi)) = \chi_{\varepsilon^{-1}Y'}(\xi)&\frac{B_1^\varepsilon g(\xi)}{1 + \lambda_1^\varepsilon(\xi)} \\ &-\widehat{w}^*(\xi) - (\chi_{\varepsilon^{-1}Y'}(\xi)B_1^\varepsilon w^*(\xi) - \widehat{w}^*(\xi)). \end{aligned}$$

According to Proposition 5.1, the last term tends to zero in $L^2(\mathbb{R}^N)$. It suffices to show

$$\chi_{\varepsilon^{-1}Y'}(\xi)\frac{B_1^\varepsilon g(\xi)}{1 + \lambda_1^\varepsilon(\xi)} - \widehat{w}^*(\xi) \to 0 \quad \text{in} \quad L^2(\mathbb{R}_\xi^N). \tag{6.4}$$

Note that w^* satisfies the homogenized equation $A^* w^* + w^* = g$ in $\mathbb{R}^N$, which is equivalent to

$$\left(\frac{1}{2}D_{k\ell}^2 \lambda_1(0)\xi_k \xi_\ell + 1\right) \widehat{w}^*(\xi) = \widehat{g}(\xi), \quad \xi \in \mathbb{R}^N.$$

So, (6.4) is reduced to

$$\chi_{\varepsilon^{-1}Y'}(\xi)\frac{B_1^\varepsilon g(\xi)}{1+\lambda_1^\varepsilon(\xi)} - \frac{\widehat{g}(\xi)}{1+\frac{1}{2}D^2_{k\ell}\lambda_1(0)\xi_k\xi_\ell} \to 0 \quad \text{in} \quad L^2(\mathbb{R}^N_\xi). \tag{6.5}$$

The above expression can be written in the form

$$\frac{a^\varepsilon + b^\varepsilon}{c^\varepsilon},$$

where

$$a^\varepsilon = \left(1 + \frac{1}{2}\ D^2_{k\ell}\lambda_1(0)\xi_k\xi_\ell\right)\left[\chi_{\varepsilon^{-1}Y'}(\xi)B_1^\varepsilon g(\xi) - \widehat{g}(\xi)\right],$$

$$b^\varepsilon = -\left(\lambda_1^\varepsilon(\xi) - \frac{1}{2}D^2_{k\ell}\lambda_1(0)\xi_k\xi_\ell\right)\widehat{g}(\xi),$$

$$c^\varepsilon = (1+\lambda_1^\varepsilon(\xi))\left(1 + \frac{1}{2}\ D^2_{k\ell}\lambda_1(0)\xi_k\xi_\ell\right).$$

Now we have the convergence

$$\frac{a^\varepsilon}{c^\varepsilon} = \frac{\chi_{\varepsilon^{-1}Y'}(\xi)B_1^\varepsilon g(\xi) - \widehat{g}(\xi)}{1+\lambda_1^\varepsilon(\xi)} \to 0 \text{ in } L^2(\mathbb{R}^N_\xi)$$

because $[1+\lambda_1^\varepsilon(\xi)] \geq 1$ and by the virtue of Proposition 5.1.

The convergence of $\frac{b^\varepsilon}{c^\varepsilon}$ is not immediate. To show this, we split the energy into three parts, taking $\gamma > 0$ as a fixed constant:

$$\int_{\substack{|\xi|\leq\delta\varepsilon^{-1}\\|\xi|\leq\gamma}} \left(\frac{b^\varepsilon}{c^\varepsilon}\right)^2 d\xi + \int_{\substack{|\xi|\leq\delta\varepsilon^{-1}\\|\xi|>\gamma}} \left(\frac{b^\varepsilon}{c^\varepsilon}\right)^2 d\xi + \int_{|\xi|>\delta\varepsilon^{-1}} \left(\frac{b^\varepsilon}{c^\varepsilon}\right)^2 d\xi.$$

In the first two parts, we use the estimate

$$\left|\lambda_1^\varepsilon(\xi) - \frac{1}{2}D^2_{k\ell}\lambda_1(0)\xi_k\xi_\ell\right| \leq c|\xi|^3\varepsilon \quad \text{for} \quad |\xi| \leq \delta\varepsilon^{-1}, \tag{6.6}$$

which holds since $\lambda_1(0) = D\lambda_1(0) = 0$ (see Proposition 1.5). In the first integral, we have $c^\varepsilon \geq 1$ and $|b^\varepsilon(\xi)| \leq c\gamma^3\varepsilon|\widehat{g}(\xi)|$, and consequently it is less than

$$c\varepsilon^2 \int_{\mathbb{R}^N} |\widehat{g}(\xi)|^2 d\xi$$

and hence converges to zero. In the second integral, we have

$$c^\varepsilon \geq \frac{1}{2}\lambda_1^\varepsilon(\xi)D^2_{k\ell}\lambda_1(0)\xi_k\xi_\ell \geq c|\xi|^4 \geq c\gamma|\xi|^3 \quad \text{since} \quad |\xi| \geq \gamma > 0.$$

With regards to b^ε, we still have $|b^\varepsilon(\xi)| \leq c|\xi|^3\varepsilon|\widehat{g}(\xi)|$, and so the second integral also converges to zero.

In the third integral, we use the bounds

$$|b^\varepsilon| \leq \left(\lambda_1^\varepsilon(\xi) + \frac{1}{2}D^2_{k\ell}\lambda_1(0)\xi_k\xi_\ell\right)|\widehat{g}(\xi)|,$$

$$c^\varepsilon \geq \lambda_1^\varepsilon(\xi) + \frac{1}{2}D^2_{k\ell}\lambda_1(0)\xi_k\xi_\ell.$$

Thus the third integral is estimated from above by

$$\int_{|\xi|>\delta\varepsilon^{-1}} |\widehat{g}(\xi)|^2 d\xi.$$

Obviously, this tends to zero as $\varepsilon \to 0$ since $g \in L^2(\mathbb{R}^N)$. □

6.2. Corrector result in $\mathbb{R}^N$. This section is devoted to the proof of Theorem 1.8 concerning the Bloch approximation θ^ε. The proof consists of several steps which correspond to estimations of the required energy in different regions in the Fourier space (in a neighborhood of the origin $|\eta| \leq \delta$ and in its complement $|\eta| > \delta$).

Step 1. We decompose u^ε as follows:

$$u^\varepsilon = v^\varepsilon + P_1^\varepsilon u^\varepsilon,$$

where v^ε and $P_1^\varepsilon u^\varepsilon$ are defined in (1.21) and (1.27), respectively. Thanks to Proposition 3.1, it is enough to prove

$$\|P_1^\varepsilon u^\varepsilon - \theta^\varepsilon\|_{L^2(\mathbb{R}^N)} \to 0, \tag{6.7}$$

$$|P_1^\varepsilon u^\varepsilon - \theta^\varepsilon|_{H^1(\mathbb{R}^N)} \leq c\varepsilon \|f\|_{L^2(\mathbb{R}^N)}. \tag{6.8}$$

Step 2. We estimate the energies in the region $|\xi| > \delta\varepsilon^{-1}$. To this end, we introduce the quantities

$$\theta^{\varepsilon,\delta}(x) = \int_{\substack{\xi\in\varepsilon^{-1}Y' \\ |\xi|>\delta\varepsilon^{-1}}} \widehat{u}^*(\xi) e^{ix\cdot\xi} \phi_1^\varepsilon(x;\xi) d\xi, \tag{6.9}$$

$$P_1^{\varepsilon,\delta} u^\varepsilon(x) = \int_{\substack{\xi\in\varepsilon^{-1}Y' \\ |\xi|>\delta\varepsilon^{-1}}} B_1^\varepsilon u^\varepsilon(\xi) e^{ix\cdot\xi} \phi_1^\varepsilon(x;\xi) d\xi. \tag{6.10}$$

We will obtain the estimates

$$\|\theta^{\varepsilon,\delta}\|_{L^2(\mathbb{R}^N)} \leq c\varepsilon \|f\|_{H^{-1}(\mathbb{R}^N)}, \tag{6.11}$$

$$|\theta^{\varepsilon,\delta}|_{H^1(\mathbb{R}^N)} \leq c\varepsilon \|f\|_{L^2(\mathbb{R}^N)}, \tag{6.12}$$

$$\|P_1^{\varepsilon,\delta} u^\varepsilon\|_{L^2(\mathbb{R}^N)} \leq c\varepsilon \|f\|_{H^{-1}(\mathbb{R}^N)}, \tag{6.13}$$

$$|P_1^{\varepsilon,\delta} u^\varepsilon|_{H^1(\mathbb{R}^N)} \leq c\varepsilon \|f\|_{L^2(\mathbb{R}^N)}. \tag{6.14}$$

We start with (6.14). Using Lemma 2.3 with $\rho = \phi_1$ and inequalities (2.5), we get

$$|P_1^{\varepsilon,\delta} u^\varepsilon|^2_{H^1(\mathbb{R}^N)} \leq c \int_{\substack{\xi\in\varepsilon^{-1}Y' \\ |\xi|>\delta\varepsilon^{-1}}} |B_1^\varepsilon u^\varepsilon(\xi)|^2 \lambda_1^\varepsilon(\xi) d\xi.$$

Now (6.14) easily follows if we use (1.22) and (1.24). Next, we prove (6.12). Following the above procedure, we get

$$|\theta^{\varepsilon,\delta}|^2_{H^1(\mathbb{R}^N)} \leq c \int_{\substack{\xi\in\varepsilon^{-1}Y' \\ |\xi|>\delta\varepsilon^{-1}}} |\widehat{u}^*(\xi)|^2 |\xi|^2 d\xi. \tag{6.15}$$

If $f \in L^2(\mathbb{R}^N)$, then it is well known that $u^* \in H^2(\mathbb{R}^N)$ and

$$\int_{\mathbb{R}^N} |\xi|^4 |\widehat{u}^*(\xi)|^2 d\xi \leq c \int_{\mathbb{R}^N} |\widehat{f}(\xi)|^2 d\xi. \tag{6.16}$$

Combining (6.15) and (6.16), we easily get (6.12). We now show (6.11). By Parseval's identity, we have

$$\|\theta^{\varepsilon,\delta}\|^2_{L^2(\mathbb{R}^N)} = \int_{\substack{\xi \in \varepsilon^{-1}Y' \\ |\xi| > \delta\varepsilon^{-1}}} |\widehat{u}^*(\xi)|^2 d\xi \leq c_\delta \varepsilon^2 \int_{\substack{\xi \in \varepsilon^{-1}Y' \\ |\xi| > \delta\varepsilon^{-1}}} |\xi|^{-2} |\widehat{f}(\xi)|^2 d\xi,$$

since u^* and f are related by the homogenized equation $A^* u^* = f$ in $\mathbb{R}^N$. This clearly implies

$$\|\theta^{\varepsilon,\delta}\|^2_{L^2(\mathbb{R}^N)} \leq c_\delta \varepsilon^2 \int_{\substack{\xi \in \varepsilon^{-1}Y' \\ |\xi| > \delta\varepsilon^{-1}}} (1 + |\xi|^2)^{-1} |\widehat{f}(\xi)|^2 d\xi = c_\delta \varepsilon^2 \|f\|^2_{H^{-1}(\mathbb{R}^N)}.$$

The proof of (6.13) is completely analogous.

Step 3. Now, we consider the energies in $|\xi| \leq \delta\varepsilon^{-1}$. To this end, let us define

$$\omega^\varepsilon(x) = \int_{|\xi| \leq \delta\varepsilon^{-1}} (B_1^\varepsilon u^\varepsilon(\xi) - \widehat{u}^*(\xi)) e^{ix\cdot\xi} \phi_1^\varepsilon(x;\xi) d\xi \tag{6.17}$$

and show that

$$\|\omega^\varepsilon\|_{L^2(\mathbb{R}^N)} \to 0, \tag{6.18}$$

$$|\omega^\varepsilon|_{H^1(\mathbb{R}^N)} \leq c\varepsilon \|f\|_{L^2(\mathbb{R}^N)}. \tag{6.19}$$

To prove (6.18), we decompose the integrand as follows:

$$B_1^\varepsilon u^\varepsilon - \widehat{u}^* = B_1^\varepsilon (u^\varepsilon - u^*) + (B_1^\varepsilon u^* - \widehat{u}^*).$$

By Parseval's equality, the first term in the L^2-norm is bounded above by $\|u^\varepsilon - u^*\|_{L^2(\mathbb{R}^N)}$ which, by our hypothesis, converges to zero. That the second term converges to zero in $L^2(\mathbb{R}^N)$ is proved in Proposition 5.1.

Next, we turn are attention to the proof of (6.19). By Lemma 2.1, we have

$$|\omega^\varepsilon|^2_{H^1(\mathbb{R}^N)} \leq c \int_{|\xi| \leq \delta\varepsilon^{-1}} \lambda_1^\varepsilon(\xi) |B_1^\varepsilon u^\varepsilon(\xi) - \widehat{u}^*(\xi)|^2 d\xi. \tag{6.20}$$

To estimate the above integral, we write the integrand as

$$B_1^\varepsilon u^\varepsilon(\xi) - \widehat{u}^*(\xi) = \lambda_1^\varepsilon(\xi)^{-1} (B_1^\varepsilon f(\xi) - \widehat{f}(\xi)) + \Big[\lambda_1^\varepsilon(\xi)^{-1} - \Big(\frac{1}{2} D^2_{k\ell} \lambda_1(0) \xi_k \, \xi_\ell\Big)^{-1}\Big] \widehat{f}(\xi).$$

Thus we get, using (1.24), that

$$\begin{cases} |\omega^\varepsilon|^2_{H^1(\mathbb{R}^N)} \;\leq\; c \displaystyle\int_{|\xi| \leq \delta\varepsilon^{-1}} \frac{|B_1^\varepsilon f(\xi) - \widehat{f}(\xi)|^2}{|\xi|^2} d\xi \\ \qquad\qquad\quad + \; c \displaystyle\int_{|\xi| \leq \delta\varepsilon^{-1}} \lambda_1^\varepsilon(\xi) \Big| \lambda_1^\varepsilon(\xi)^{-1} - \Big(\frac{1}{2} D^2_{k\ell} \lambda_1(0) \xi_k \xi_\ell\Big)^{-1} \Big|^2 |\widehat{f}(\xi)|^2 d\xi. \end{cases} \tag{6.21}$$

To estimate the first term on the right-hand side of (6.21), we represent the integrand as

$$\frac{B_1^\varepsilon f(\xi) - \widehat{f}(\xi)}{|\xi|} = \int_{\mathbb{R}^N} f(x) e^{-ix\cdot\xi} \frac{(\phi_1^\varepsilon(x;\xi) - \phi_1^\varepsilon(x;0))}{|\xi|} dx.$$

Applying Lemma 2.5 and using $\|\phi_1(\cdot;\eta) - \phi_1(\cdot;0)\|_{L^2(Y)} \leq c|\eta|$ for $|\eta| \leq \delta$, we get

$$\int_{|\xi|\leq\delta\varepsilon^{-1}} \frac{|B_1^\varepsilon f(\xi) - \widehat{f}(\xi)|^2}{|\xi|^2} d\xi \leq c\varepsilon^2 \|f\|^2_{L^2(\mathbb{R}^N)}.$$

The second term on the right-hand side of (6.21) can be rewritten, using the homogenized equation, as

$$\int_{|\xi|\leq\delta\varepsilon^{-1}} \frac{|\lambda_1^\varepsilon(\xi) - \frac{1}{2}D^2_{k\ell}\lambda_1(0)\xi_k\xi_\ell|^2}{\lambda_1^\varepsilon(\xi)} |\widehat{u}^*(\xi)|^2 d\xi.$$

Using (6.6) and (1.24), we see that the above integral is estimated from above by

$$c\varepsilon^2 \int_{|\xi|\leq\delta\varepsilon^{-1}} |\xi|^4 |\widehat{u}^*(\xi)|^2 d\xi \leq c\varepsilon^2 \|f\|^2_{L^2(\mathbb{R}^N)}.$$

This establishes (6.19) and hence the result. □

6.3. Asymptotic expansion of the Bloch approximation. In this concluding section, we prove Theorem 1.11.

Proof of (i). We have the following decomposition:

$$\theta^\varepsilon(x) - u^*(x) = z^\varepsilon(x) + \theta^{\varepsilon,\delta}(x) + u^{*,\delta}(x), \tag{6.22}$$

where

$$z^\varepsilon(x) = \int_{\substack{\xi\in\varepsilon^{-1}Y' \\ |\xi|\leq\delta\varepsilon^{-1}}} \widehat{u}^*(\xi) e^{ix\cdot\xi} (\phi_1^\varepsilon(x;\xi) - \phi_1^\varepsilon(x;0)) d\xi, \tag{6.23}$$

$$u^{*,\delta}(x) = \frac{1}{(2\pi)^{N/2}} \int_{|\xi|>\delta\varepsilon^{-1}} \widehat{u}^*(\xi) e^{ix\cdot\xi} d\xi, \tag{6.24}$$

and $\theta^{\varepsilon,\delta}$ is defined in (6.9).

The second term has already been estimated in the L^2-norm (see (6.11)). The same proof shows that the third term admits a bound

$$\|u^{*,\delta}\|_{L^2(\mathbb{R}^N)} \leq c\varepsilon \|f\|_{H^{-1}(\mathbb{R}^N)} \leq c\varepsilon \|u^*\|_{H^1(\mathbb{R}^N)}. \tag{6.25}$$

To estimate the first term on the right-hand side of (6.22), we must proceed differently. In fact, it is essential to use Lemma 2.3. We see then that

$$\|z^\varepsilon\|^2_{L^2(\mathbb{R}^N)} = \int_{\substack{\xi\in\varepsilon^{-1}Y' \\ |\xi|\leq\delta\varepsilon^{-1}}} |\widehat{u}^*(\xi)|^2 \|\phi_1^\varepsilon(\cdot;\xi) - \phi_1^\varepsilon(\cdot;0)\|^2_{L^2(Y)} d\xi.$$

Using the Lipschitz continuity of the map $\eta \mapsto \phi_1(\cdot;\eta) \in L^2(Y)$ for $|\eta| \leq \delta$, we see that the above integral can be majorized, and we obtain

$$\|z^\varepsilon\|^2_{L^2(\mathbb{R}^N)} \leq c\varepsilon^2 \int_{|\xi|\leq\delta\varepsilon^{-1}} |\widehat{u}^*(\xi)|^2 |\xi|^2 d\xi \leq c\varepsilon^2 |u^*|^2_{H^1(\mathbb{R}^N)}. \tag{6.26}$$

This finishes the proof of (i). We note that we cannot, in general, assert that

$$|u^*|_{H^1(\mathbb{R}^N)} \leq c\|f\|_{H^{-1}(\mathbb{R}^N)}$$

as we are working on the entire space $\mathbb{R}^N$.

Proof of (ii). Because of (i), it suffices to prove

$$\left|\theta^\varepsilon - u^* - \varepsilon\chi_k^\varepsilon \frac{\partial u^*}{\partial x_k}\right|_{H^1(\mathbb{R}^N)} \leq c\varepsilon\|f\|_{L^2(\mathbb{R}^N)}. \tag{6.27}$$

To this end, we use once again the decomposition (6.22) for $(\theta^\varepsilon - u^*)$ in terms of z^ε, $\theta^{\varepsilon,\delta}$, and $u^{*,\delta}$. For $\theta^{\varepsilon,\delta}$, we have the estimate (6.12). For $u^{*,\delta}$, we can easily derive the estimate

$$|u^{*,\delta}|^2_{H^1(\mathbb{R}^N)} \leq c\int_{|\xi|>\delta\varepsilon^{-1}} |\xi|^2|\widehat{u}^*(\xi)|^2 d\xi \leq c_\delta\varepsilon^2\|f\|^2_{L^2(\mathbb{R}^N)}. \tag{6.28}$$

Thus, we are reduced to obtaining the estimate

$$\left|z^\varepsilon - \varepsilon\chi_k^\varepsilon \frac{\partial u^*}{\partial x_k}\right|_{H^1(\mathbb{R}^N)} \leq c\varepsilon\|f\|_{L^2(\mathbb{R}^N)}. \tag{6.29}$$

To this end, we use the representation

$$\frac{\partial u^*}{\partial x_k}(x) = \frac{1}{(2\pi)^{N/2}}\int_{\mathbb{R}^N}(i\xi_k)\widehat{u}^*(\xi)e^{ix\cdot\xi}d\xi,$$

and combine it with the representation (6.23) for z^ε. We get

$$\begin{aligned} z^\varepsilon(x) - \varepsilon\chi_k^\varepsilon(x)\frac{\partial u^*}{\partial x_k}(x) = &\int_{|\xi|\leq\delta\varepsilon^{-1}} \widehat{u}^*(\xi)e^{ix\cdot\xi}\left(\phi_1^\varepsilon(x;\xi) - \phi_1^\varepsilon(x;0) - ip^{(0)}\chi_k^\varepsilon(x)\varepsilon\xi_k\right)d\xi \\ &-\int_{|\xi|>\delta\varepsilon^{-1}} ip^{(0)}\chi_k^\varepsilon(x)\varepsilon\xi_k\widehat{u}^*(\xi)e^{ix\cdot\xi}d\xi. \end{aligned} \tag{6.30}$$

To estimate the first term on the right-hand side of (6.30), we appeal to Lemma 2.3. Further, we use

$$\left\|\phi_1(\cdot;\eta) - \phi_1(\cdot;0) - ip^{(0)}\chi_k(\cdot)\eta_k\right\|_{H^1(Y)} \leq c|\eta|^2 \quad \text{for} \quad |\eta|\leq\delta. \tag{6.31}$$

The estimate on the second term on the right-hand side of (6.30) is more straightforward. We finally get

$$\left|z^\varepsilon - \varepsilon\chi_k^\varepsilon\frac{\partial u^*}{\partial x_k}\right|^2_{H^1(\mathbb{R}^N)} \leq c\varepsilon^2\int_{\mathbb{R}^N}|\xi|^4|\widehat{u}^*(\xi)|^2 d\xi.$$

This completes the proof of (6.29) and hence (ii).

Proof of (iii). Consider again the decomposition (6.22). Thanks to (6.9) and (6.15), we have the estimates

$$\|\theta^{\varepsilon,\delta}\|_{L^2(\mathbb{R}^N)} \leq c\varepsilon^2\|f\|_{L^2(\mathbb{R}^N)}, \tag{6.32}$$

$$|\theta^{\varepsilon,\delta}|_{H^1(\mathbb{R}^N)} \leq c\varepsilon^2|f|_{H^1(\mathbb{R}^N)}. \tag{6.33}$$

Similar techniques imply

$$\|u^{*,\delta}\|_{L^2(\mathbb{R}^N)} \le c\varepsilon^2 \|f\|_{L^2(\mathbb{R}^N)}, \tag{6.34}$$

$$|u^{*,\delta}|_{H^1(\mathbb{R}^N)} \le c\varepsilon^2 |f|_{H^1(\mathbb{R}^N)}. \tag{6.35}$$

On the other hand, it is clear from the representation (6.30) that

$$\Big\| z^\varepsilon - \varepsilon \chi^\varepsilon_k \frac{\partial u^*}{\partial x_k} \Big\|_{L^2(\mathbb{R}^N)} \le c\varepsilon^2 \|f\|_{L^2(\mathbb{R}^N)}. \tag{6.36}$$

Thus, it is enough to obtain the estimate

$$\Big| \theta^\varepsilon - u^* - \varepsilon \chi^\varepsilon_k \frac{\partial u^*}{\partial x_k} + \varepsilon^2 (\chi^\varepsilon_{k\ell} + \beta^{(2)}_{k\ell}) \frac{\partial^2 u^*}{\partial x_k \partial x_\ell} \Big|_{H^1(\mathbb{R}^N)} \le c\varepsilon^2 |f|_{H^1(\mathbb{R}^N)}. \tag{6.37}$$

Thanks to (6.33) and (6.35), we are reduced to showing that

$$\Big| z^\varepsilon - \varepsilon \chi^\varepsilon_k \frac{\partial u^*}{\partial x_k} + \varepsilon^2 (\chi^\varepsilon_{k\ell} + \beta^{(2)}_{k\ell}) \frac{\partial^2 u^*}{\partial x_k \partial x_\ell} \Big|_{H^1(\mathbb{R}^N)} \le c\varepsilon^2 |f|_{H^1(\mathbb{R}^N)}. \tag{6.38}$$

We can write

$$\begin{aligned} z^\varepsilon(x) - \varepsilon \chi^\varepsilon_k(x) \frac{\partial u^*}{\partial x_k}(x) &+ \varepsilon^2 (\chi^\varepsilon_{k\ell}(x) + \beta^{(2)}_{k\ell}) \frac{\partial^2 u^*}{\partial x_k \partial x_\ell}(x) \\ &= \int_{|\xi| \le \delta\varepsilon^{-1}} \widehat{u}^*(\xi) e^{ix\cdot\xi} [\phi^\varepsilon_1(x;\xi) - \phi^\varepsilon_1(x;0) - ip^{(0)} \chi^\varepsilon_k(x) \varepsilon \xi_k i \\ &\qquad + p^{(0)} (\chi^\varepsilon_{k\ell}(x) + \beta^{(2)}_{k\ell}) \varepsilon^2 \xi_k \xi_\ell] d\xi \\ &\quad - \int_{|\xi| > \delta\varepsilon^{-1}} ip^{(0)} \chi^\varepsilon_k(x) \varepsilon \xi_k \widehat{u}^*(\xi) e^{ix\cdot\xi} d\xi \\ &\quad + \int_{|\xi| > \delta\varepsilon^{-1}} p^{(0)} (\chi^\varepsilon_{k\ell}(x) + \beta^{(2)}_{k\ell}) \varepsilon^2 \xi_k \xi_\ell \widehat{u}^*(\xi) e^{ix\cdot\xi} d\xi. \end{aligned} \tag{6.39}$$

The analysis of the right-hand side of (6.39) is similar to that of (6.30). The new information needed is the following:

$$\Big\| \phi_1(\cdot;\eta) - \phi_1(\cdot;0) - ip^{(0)} \chi_k(\cdot) \eta_k + p^{(0)} (\chi_{k\ell}(\cdot) + \beta^{(2)}_{k\ell}) \eta_k \eta_\ell \Big\|_{H^1(Y)} \le c|\eta|^3 \tag{6.40}$$

for $|\eta| \le \delta$, which is a simple consequence of Proposition 1.10. The proof is concluded via a simple application of Lemma 2.3. □

Acknowledgments. This research stems from a couple of visits by R. Orive and M. Vanninathan to the Center of Mathematical Modelling and to the Department of Mathematical Engineering of the University of Chile, supported by the FONDAP's Programme on Mathematical-Mechanics. These authors warmly thank these institutions.

REFERENCES

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[2] G. Allaire and C. Conca, *Bloch wave homogenization for a spectral problem in fluid-solid structures*, Arch. Ration. Mech. Anal., 135 (1996), pp. 197–257.

[3] G. Allaire and C. Conca, *Boundary layers in the homogenization of a spectral problem in fluid-solid structures*, SIAM J. Math. Anal., 29 (1998), pp. 343–379.

[4] M. Avellaneda, L. Berlyand, and J.-F. Clouet, *Frequency-dependent acoustics of composites with interfaces*, SIAM J. Appl. Math., 60 (2000), pp. 2143–2181.

[5] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.

[6] F. Bloch, *Über die quantenmechanik der electronen in kristallgittern*, Z. Phys., 52 (1928), pp. 555–600.

[7] H. Brezis, *Analyse Fonctionnelle, Théorie et Applications*, Collection Mathématiques Appliquées pour la Maîtrise, Masson, Paris, 1983.

[8] C. Conca, *Bloch waves*, in Encyclopaedia Mathematics, M. Hazewinkel et al., eds., Kluwer Academic Publishers, Amsterdam, 1999, pp. 72–74.

[9] C. Conca, R. Orive, and M. Vanninathan, *Bloch Approximation in Bounded Domains*, manuscript, 2001.

[10] C. Conca, J. Planchard, and M. Vanninathan, *Fluids and Periodic Structures*, Rech. Math. Appl., 38, John Wiley/Masson, New York, Paris, 1995.

[11] C. Conca and M. Vanninathan, *Homogenization of periodic structures via Bloch decomposition*, SIAM J. Appl. Math., 57 (1997), pp. 1639–1659.

[12] C. Conca and M. Vanninathan, *On uniform H^2-estimates in periodic homogenization*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 499–517.

[13] G. Floquet, *Sur les équations différentielles linéaires à coefficients périodiques*, Ann. Ecole Norm. Sér. 2, 12 (1883), pp. 47–89.

[14] P. Gérard, P.-A. Markowich, N.-J. Mauser, and F. Poupaud, *Homogenization limits and Wigner transforms*, Comm. Pure Appl. Math., 50 (1997), pp. 323–379.

[15] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.

[16] F. Murat and L. Tartar, *H-convergence*, in Topics in the Mathematical Modeling of Composite Materials, Progr. Nonlinear Differential Equations Appl. 31, A. Cherkaev and R. Kohn, eds., Birkhäuser, Boston, 1997, pp. 21–44.

[17] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.

[18] F. Odeh and J. Keller, *Partial differential equations with periodic coefficients and Bloch waves in crystals*, J. Math. Phys., 5 (1964), pp. 1499–1504.

[19] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Academic Press, New York, 1978.

[20] F. Santosa and W. W. Symes, *A dispersive effective medium for wave propagation in periodic composites*, SIAM J. Appl. Math., 51 (1991), pp. 984–1005.

[21] L. Tartar, *Problèmes d'homogénéisation dans les équations aux dérivées partielles*, in Cours Peccot au Collège de France, 1977.

[22] C. Wilcox, *Theory of Bloch waves*, J. Anal. Math., 33 (1978), pp. 146–167.

SIAM J. MATH. ANAL.
Vol. 33, No. 5, pp. 1199–1210

ON THE HARDY–LITTLEWOOD THEOREM FOR FUNCTIONS OF BOUNDED VARIATION[*]

PAULO R. ZINGANO[†] AND STANLY L. STEINBERG[‡]

Abstract. We derive here an improved version of the well-known Hardy–Littlewood theorem for functions of bounded variation in the classical sense of Jordan. Results for the positive and negative variations are also discussed.

Key words. functions of bounded variation, Hardy–Littlewood theorem, essential bounded variation, positive and negative variations, absolutely continuous functions

AMS subject classifications. Primary, 26A45; Secondary, 26A46

PII. S0036141000369307

1. Introduction. In this paper we will point out a fundamental property of functions of bounded variation (in the classical sense of C. Jordan [15], [16]) which has apparently not been noticed in the well-developed theory of such functions. We recall that $f : [a, b] \to R$ is of bounded (or finite) variation on the interval $[a, b]$, written $f \in \mathrm{BV}(a, b)$, if its total variation given by

$$\mathrm{TV}[f; a, b] = \sup_{\pi} \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})| \tag{1.1}$$

is finite, where the supremum is taken over all possible partitions $\pi = \{x_0 < x_1 < \cdots < x_n\}$ of $[a, b]$. Given $f \in \mathrm{BV}(a, b)$, its positive and negative variations are similarly defined by

$$\mathrm{PV}[f; a, b] = \sup_{\pi} \sum_{i=1}^{n} \big(f(x_i) - f(x_{i-1}) \big)_+, \tag{1.2a}$$

$$\mathrm{NV}[f; a, b] = \sup_{\pi} \sum_{i=1}^{n} \big(f(x_i) - f(x_{i-1}) \big)_-, \tag{1.2b}$$

where, for an arbitrary $\theta \in R$, θ_+ and θ_- denote the positive and negative parts of θ, i.e.,

$$\theta_+ = \frac{1}{2}(|\theta| + \theta), \qquad \theta_- = \frac{1}{2}(|\theta| - \theta). \tag{1.3}$$

The concept of bounded variation plays a key role in many important topics in the theory of functions [14], [22], [35], measure and integration [5], [6], [14], [26], partial differential equations [10], [20], [28], [29], numerical analysis [10], [11], [19], applied

[*] Received by the editors March 22, 2000; accepted for publication (in revised form) October 11, 2001; published electronically April 8, 2002.
http://www.siam.org/journals/sima/33-5/36930.html

[†] Departamento de Matemática Pura e Aplicada, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS 91509, Brazil (zingano@mat.ufrgs.br). This work was partly supported by CNPq, Brazil, and Cray Research, Inc.

[‡] Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131 (stanly@math.unm.edu). This work was partly supported by NSF grant CCR-9531828 and AFOSR grant 98-0015.

mathematics [3], [18], [27], and other disciplines. It has been generalized with many fruitful applications to higher dimensions [9], [14], [20], [28], [33], [34], functions with values in normed or metric spaces and set-valued functions [1], [23], [24], [26], and more general notions such as pth-variations [31], [32], harmonic and Λ-variations [30], Φ-variations [21], [30], [32], and other classes of functions; see, e.g., [2], [7], and the recent literature. In this respect, e.g., we should observe that, taking into account [5, section 2.9], the results of the present paper can be carried over to mappings of bounded variation with values in a reflexive Banach space.

Our concern here is certain equivalent characterizations for the space $\mathrm{BV}(a,b)$. It is well known that any $f \in \mathrm{BV}(a,b)$ is bounded and has finite side limits $f(x^-)$, $f(x^+)$ at every $x \in \,]\, a,\, b\,[$; also, $f(a^+)$, $f(b^-)$ are well defined so that all discontinuities of f (if any) are simple and, by the finiteness of (1.1), are at most denumerable. Changing the value of $f \in \mathrm{BV}(a,b)$ at a single point may change its total variation; by the remarks just made, it is natural to assume that f has no external saltus, i.e., one has $f(a) = f(a^+)$, $f(b) = f(b^-)$, and

$$f(x) \;\in\; [\![\, f(x^-)\,,\, f(x^+)\,]\!] \qquad \forall\, x \,\in\,]\, a,\, b\,[, \tag{1.4}$$

where, for real α , β, $[\![\, \alpha, \beta\,]\!]$ denotes the smallest closed interval containing the points α, β, that is, $[\![\, \alpha, \beta\,]\!] = [\, \min(\alpha,\beta), \max(\alpha,\beta)\,]$. Following [22], we call $f \in \mathrm{BV}(a,b)$ *normal* if it has no external saltus on $[\, a,\, b\,]$. It is easy to see that if f, $g \in \mathrm{BV}(a,b)$ are equal a.e. and f is normal in the sense just given, then

$$\mathrm{TV}\,[\, f\,;\, a,b\,] \;\leq\; \mathrm{TV}\,[\, g\,;\, a,b\,] \tag{1.5}$$

and similarly for the positive and negative variations; moreover, if g is also normal on $[\, a,b\,]$, then these quantities are the same. This motivates the following definition: given an arbitrary f in the space $L^1(a,b)$ of Lebesgue measurable functions which are integrable on $[\, a,\, b\,]$, we say that f is of (essential) bounded variation on $[\, a,\, b\,]$ when there exists some $g \in \mathrm{BV}(a,b)$ such that $f = g$ a.e., and, modifying g if necessary so as to make g normal, the quantities $\mathrm{TV}\,[\, g\,;\, a,b\,]$, $\mathrm{PV}\,[\, g\,;\, a,b\,]$, $\mathrm{NV}\,[\, g\,;\, a,b\,]$ are said to be the *essential variations* of f, so that

$$\mathrm{TV}_{\mathrm{ess}}\,[\, f\,;\, a,b\,] \;=\; \mathrm{TV}\,[\, g\,;\, a,b\,] \tag{1.6}$$

when $f \;=\; g$ a.e. with $g \;\in\; \mathrm{BV}(a,b)$ normal; likewise, one has $\mathrm{PV}_{\mathrm{ess}}\,[\, f\,;\, a,b\,] \;=\; \mathrm{PV}\,[\, g\,;\, a,b\,]$ and $\mathrm{NV}_{\mathrm{ess}}\,[\, f\,;\, a,b\,] \;=\; \mathrm{NV}\,[\, g\,;\, a,b\,]$ in this case. This is equivalent to setting

$$\mathrm{TV}_{\mathrm{ess}}\,[\, f\,;\, a,b\,] \;=\; \inf_{\substack{u \,\in\, \mathrm{BV}(a,b) \\ u \,=\, f \;\; a.e.}} \mathrm{TV}\,[\, u\,;\, a,b\,] \tag{1.7}$$

and similarly for the essential positive and negative variations; other slightly different (but equivalent) characterizations are given in [2], [4], [12], [23]. Clearly, (1.6), (1.7) provide the appropriate notions when we do not want to distinguish functions which are equal a.e., as in the case of L^p spaces. In particular, it is convenient for the following well-known result. *Given $f \in L^1(a,b)$, the statements below are equivalent* *to one another:*

$$\mathrm{TV}_{\mathrm{ess}}\,[\,f\,;\,a,b\,]\;=\;C, \tag{1.8a}$$

$$\sup_{h>0}\;\frac{1}{h}\int_a^{b-h}|\,f(x+h)-f(x)\,|\,dx\;=\;C, \tag{1.8b}$$

$$\limsup_{h\to 0^+}\;\frac{1}{h}\int_a^{b-h}|\,f(x+h)-f(x)\,|\,dx\;=\;C, \tag{1.8c}$$

where C stands for a nonnegative quantity (possibly infinite). Other equivalent assertions can be found in, e.g., [4], [9], [20], [25], [26], [28], [34], but there will be no need of them here. The equivalence of (1.8a) and (1.8c) is the essence of the Hardy–Littlewood theorem [13]; see also [2], [23], [35], and section 3 below. It readily gives (1.8b) once we observe that

$$\frac{1}{h}\int_a^{b-h}|\,f(x+h)-f(x)\,|\,dx\;\leq\;\mathrm{TV}_{\mathrm{ess}}\,[\,f\,;\,a,b\,]\qquad\forall\;h>0. \tag{1.9}$$

We now state our main result: when C is finite, another statement equivalent to (1.8a), (1.8c) above is that

$$\lim_{h\to 0^+}\;\frac{1}{h}\int_a^{b-h}|\,f(x+h)-f(x)\,|\,dx\;=\;C. \tag{1.10}$$

In other words, given $f\in L^1(a,b)$, *whenever the limit superior given in* (1.8c) *is finite, the corresponding limit inferior has necessarily the same value*, so that the limit (1.10) is well defined. This will be established in section 3, along with other results which come naturally in the analysis. When f is absolutely continuous on $[\,a,b\,]$, these properties are easily obtained (see, e.g., [5], [25], [26]), and we find it convenient to review this case first, which forms the subject of section 2.

2. Absolutely continuous functions. We start the analysis with an important subclass of the space $\mathrm{BV}(a,b)$, namely, the space $AC(a,b)$ of absolutely continuous functions on $[\,a,b\,]$, i.e., integrals of functions in $L^1(a,b)$. These functions share a number of properties which are well known for mappings in $C^1([\,a,b\,])$, the space of continuously differentiable functions, as the following result illustrates. One should recall that an (arbitrary) function $\varphi\in\mathrm{BV}(a,b)$ of bounded variation has a finite derivative at almost every point in the interval concerned; moreover, its derivative φ' belongs to $L^1(a,b)$; see, e.g., [5], [6], [14], [17].

THEOREM 2.1. *Let $\varphi\in\mathrm{AC}(a,b)$. Then $\varphi\in\mathrm{BV}(a,b)$ and its total, positive, and negative variations on $[\,a,b\,]$* are given by

$$\mathrm{TV}\,[\,\varphi\,;\,a,b\,]\;=\;\int_a^b|\,\varphi'(x)\,|\,dx, \tag{2.1a}$$

$$\mathrm{PV}\,[\,\varphi\,;\,a,b\,]\;=\;\int_a^b\big(\,\varphi'(x)\,\big)_+\,dx, \tag{2.1b}$$

$$\mathrm{NV}\,[\,\varphi\,;\,a,b\,]\;=\;\int_a^b\big(\,\varphi'(x)\,\big)_-\,dx, \tag{2.1c}$$

where $\varphi' \in L^1(a,b)$ is the a.e. derivative of φ, and $(\cdot)_+$, $(\cdot)_-$ denote the positive and negative parts of $(\cdot)$; cf. (1.3).

Proof. For any partition $\{ x_0 < x_1 < \cdots < x_n \}$ of $[a,b]$, we have

$$\sum_{i=1}^{n} | \varphi(x_i) - \varphi(x_{i-1}) | = \sum_{i=1}^{n} \left| \int_{x_{i-1}}^{x_i} \varphi'(x)\, dx \right| \le \int_a^b | \varphi'(x) | \, dx$$

so that $\varphi \in \mathrm{BV}(a,b)$ and

$$\mathrm{TV}[\varphi;\, a,b] \le \int_a^b |\varphi'(x)|\, dx. \tag{2.2}$$

Similarly, using (1.3), we obtain

$$\mathrm{PV}[\varphi;\, a,b] \le \int_a^b \big(\varphi'(x)\big)_+ \, dx, \qquad \mathrm{NV}[\varphi;\, a,b] \le \int_a^b \big(\varphi'(x)\big)_- \, dx.$$

To show that equality holds, let (ψ_ℓ) be a sequence of smooth (C^1) functions compactly supported in $]\,a, b\,[$ (i.e., $\psi_\ell \in C_0^1(a,b)$) and such that

$$\| \psi_\ell - \varphi' \|_{L^1(a,b)} \to 0 \quad \text{as} \quad \ell \to \infty.$$

For each ℓ, we set

$$\varphi_\ell(x) = \varphi(a) + \int_a^x \psi_\ell(s)\, ds, \quad x \in [a,b].$$

Now, given $\varepsilon > 0$, let $L(\varepsilon) > 0$ be such that $\| \psi_\ell - \varphi' \|_{L^1(a,b)} \le \frac{\varepsilon}{2}$ for all $\ell \ge L(\varepsilon)$ and, for each ℓ, let $a = x_0^{(\ell)} < x_1^{(\ell)} < \cdots < x_{n_\ell}^{(\ell)} = b$ be chosen so that

$$\sum_{i=1}^{n_\ell} | \varphi_\ell(x_i^{(\ell)}) - \varphi_\ell(x_{i-1}^{(\ell)}) | \ge \mathrm{TV}[\varphi_\ell;\, a,b] - \frac{\varepsilon}{2}.$$

Then, for each $\ell \ge L(\varepsilon)$, we have

$$\begin{aligned}
\mathrm{TV}[\varphi_\ell;\, a,b] - \frac{\varepsilon}{2} &\le \sum_{i=1}^{n_\ell} \left| \int_{x_{i-1}^{(\ell)}}^{x_i^{(\ell)}} \varphi_\ell'(s)\, ds \right| \\
&\le \sum_{i=1}^{n_\ell} \left| \int_{x_{i-1}^{(\ell)}}^{x_i^{(\ell)}} \varphi'(s)\, ds \right| + \sum_{i=1}^{n_\ell} \left| \int_{x_{i-1}^{(\ell)}}^{x_i^{(\ell)}} \big(\psi_\ell(s) - \varphi'(s)\big)\, ds \right| \\
&\le \sum_{i=1}^{n_\ell} | \varphi(x_i^{(\ell)}) - \varphi(x_{i-1}^{(\ell)}) | + \sum_{i=1}^{n_\ell} \int_{x_{i-1}^{(\ell)}}^{x_i^{(\ell)}} | \psi_\ell(s) - \varphi'(s) |\, ds \\
&\le \mathrm{TV}[\varphi;\, a,b] + \| \psi_\ell - \varphi' \|_{L^1(a,b)}
\end{aligned}$$

so that

$$\mathrm{TV}\,[\,\varphi_\ell\,;\,a,b\,]\;-\;\frac{\varepsilon}{2}\;\le\;\mathrm{TV}\,[\,\varphi\,;\,a,b\,]\;+\;\|\,\psi_\ell\;-\;\varphi'\,\|_{L^1(a,b)}.$$

Hence, for every $\ell \ge L(\varepsilon)$, we have

$$\mathrm{TV}\,[\,\varphi\,;\,a,b\,]\;\ge\;\mathrm{TV}\,[\,\varphi_\ell\,;\,a,b\,]\;-\;\varepsilon, \tag{2.3}$$

and, because

$$\mathrm{TV}\,[\,\varphi_\ell\,;\,a,b\,]\to\int_a^b|\,\varphi'(s)\,|\,ds$$

as $\ell\to\infty$, in view of the fact that, due to the smoothness of φ_ℓ,

$$\mathrm{TV}\,[\,\varphi_\ell\,;\,a,b\,]\;=\;\int_a^b|\,\varphi_\ell'(s)\,|\,ds\;=\;\int_a^b|\,\psi_\ell(s)\,|\,ds,$$

we then obtain from (2.3) that

$$\mathrm{TV}\,[\,\varphi\,;\,a,b\,]\;\ge\;\int_a^b|\,\varphi'(s)\,|\,ds\;-\;\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, this gives (2.1a) if we recall (2.2). Finally, using (1.3), we derive (2.1b) and (2.1c) in an entirely similar way. □

It is interesting to note that the converse to Theorem 2.1 is also true, that is, given $\varphi \in \mathrm{BV}(a,b)$ such that (2.1a) holds, then φ must be absolutely continuous on $[\,a,b]$; see [5, p. 165]. To complete this section, we will now derive the property (1.10) for functions in $AC(a,b)$. Clearly, it will be sufficient to establish the following result.

Theorem 2.2. *Let* $\varphi \in \mathrm{AC}(a, b+h_0)$ *for some* $h_0 > 0$. *Then*

$$\mathrm{TV}\,[\,\varphi\,;\,a,b\,]\;=\;\lim_{h\to 0^+}\;\frac{1}{h}\int_a^b\big|\,\varphi(x+h)\;-\;\varphi(x)\,\big|\,dx, \tag{2.4a}$$

$$\mathrm{PV}\,[\,\varphi\,;\,a,b\,]\;=\;\lim_{h\to 0^+}\;\frac{1}{h}\int_a^b\big(\,\varphi(x+h)\;-\;\varphi(x)\,\big)_+\,dx, \tag{2.4b}$$

$$\mathrm{NV}\,[\,\varphi\,;\,a,b\,]\;=\;\lim_{h\to 0^+}\;\frac{1}{h}\int_a^b\big(\,\varphi(x+h)\;-\;\varphi(x)\,\big)_-\,dx. \tag{2.4c}$$

Proof. By Fatou's lemma [5], [6], we have

$$\liminf_{h\to 0^+}\;\frac{1}{h}\int_a^b\big|\,\varphi(x+h)\;-\;\varphi(x)\,\big|\,dx\;\ge\;\int_a^b|\,\varphi'(x)\,|\,dx$$

so that, by (2.1a) of the previous result, we obtain

$$\liminf_{h\to 0^+}\;\frac{1}{h}\int_a^b\big|\,\varphi(x+h)\;-\;\varphi(x)\,\big|\,dx\;\ge\;\mathrm{TV}\,[\,\varphi\,;\,a,b\,]. \tag{2.5}$$

On the other hand, since

$$\begin{aligned}\int_a^b \big|\,\varphi(x+h)\,-\,\varphi(x)\,\big|\,dx \;&\le\; \int_a^b \mathrm{TV}\,[\,\varphi\,;\,x,x+h\,]\,dx\\ &=\; \int_a^b \big(\,\mathrm{TV}\,[\,\varphi\,;\,a,x+h\,]\,-\,\mathrm{TV}\,[\,\varphi\,;\,a,x\,]\,\big)\,dx\\ &=\; \int_b^{b+h} \mathrm{TV}\,[\,\varphi\,;\,a,x\,]\,dx\,-\,\int_a^{a+h} \mathrm{TV}\,[\,\varphi\,;\,a,x\,]\,dx,\end{aligned}$$

we obtain

$$\int_a^b \big|\,\varphi(x+h)\,-\,\varphi(x)\,\big|\,dx \;\le\; \int_b^{b+h} \mathrm{TV}\,[\,\varphi\,;\,a,x\,]\,dx$$

so that

$$\limsup_{h\,\to\,0^+}\;\frac{1}{h}\int_a^b \big|\,\varphi(x+h)\,-\,\varphi(x)\,\big|\,dx \;\le\; \mathrm{TV}\,[\,\varphi\,;\,a,b\,]. \tag{2.6}$$

This, together with (2.5) above, gives (2.4a). Recalling (1.3), we obtain (2.4b), (2.4c) in a completely similar way. □

In what follows, we will show that (2.4a), (2.4c) are also valid for (normal) functions of bounded variation on some interval $[\,a,\,b+h_0\,]$ containing $[\,a,b\,]$. Note that this clearly gives (1.10) for functions in $\mathrm{BV}(a,b)$, as stated in the introduction.

3. Functions with bounded variation: General case. We now turn to the main results. Let $f\in\mathrm{BV}(a,b+h_0)$ for some $h_0>0$ be given, which we assume to be normal on every point of $[\,a,b\,]$, and let $\mathrm{TV}\,[\,f\,;\,a,b+\,]$ denote the right-side limit

$$\mathrm{TV}\,[\,f\,;\,a,b+\,]\;=\;\lim_{\varepsilon\,\to\,0^+}\mathrm{TV}\,[\,f\,;\,a,b+\varepsilon\,], \tag{3.1}$$

and similarly for $\mathrm{PV}\,[\,f\,;\,a,b+\,]$ and $\mathrm{NV}\,[\,f\,;\,a,b+\,]$; these limiting quantities are simply $\mathrm{TV}\,[\,f\,;\,a,b\,]$, $\mathrm{PV}\,[\,f\,;\,a,b\,]$, and $\mathrm{NV}\,[\,f\,;\,a,b\,]$, respectively, when f is right-continuous at the point b. With this notation, we now give the following generalization of Theorem 2.2 above.

THEOREM 3.1. *Let $f\in\mathrm{BV}(a,b+h_0)$ for some $h_0>0$ with f normal on $[\,a,b\,]$. Then*

$$\lim_{h\,\to\,0^+}\;\frac{1}{h}\int_a^b \big|\,f(x+h)\,-\,f(x)\,\big|\,dx \;=\; \mathrm{TV}\,[\,f\,;\,a,b+\,], \tag{3.2a}$$

$$\lim_{h\,\to\,0^+}\;\frac{1}{h}\int_a^b \big(\,f(x+h)\,-\,f(x)\,\big)_+\,dx \;=\; \mathrm{PV}\,[\,f\,;\,a,b+\,], \tag{3.2b}$$

$$\lim_{h\,\to\,0^+}\;\frac{1}{h}\int_a^b \big(\,f(x+h)\,-\,f(x)\,\big)_-\,dx \;=\; \mathrm{NV}\,[\,f\,;\,a,b+\,]. \tag{3.2c}$$

Proof. Starting with (3.2a), we have, as observed above,

$$\int_a^b |\,f(x+h)-f(x)\,|\,dx \;\le\; h\,\mathrm{TV}\,[\,f\,;\,a,b+h\,]$$

for all $h > 0$, $h < h_0$, so that we obtain

$$\limsup_{h \to 0^+} \frac{1}{h} \int_a^b |\, f(x+h) - f(x)\,|\, dx \;\leq\; \mathrm{TV}\,[\, f\,;\, a, b+]. \tag{3.3}$$

To finish the argument, it remains to show that

$$\liminf_{h \to 0^+} \frac{1}{h} \int_a^b \big|\, f(x+h) - f(x)\,\big|\, dx \;\geq\; \mathrm{TV}\,[\, f\,;\, a, b+]. \tag{3.4}$$

For convenience, let us assume from now on that f is right-continuous at b, since setting $f(b) = f(b^+)$ does not change the values of the integrals above or the normality of f on $[\,a, b\,]$. We may then proceed as follows. Given $h > 0$, $h < h_0$, set $a_j^h \equiv a + jh$ for each $j = 0, 1, 2, \ldots$, and let $J = J_h$ be the value of j such that $a_{J-1}^h < b$ and $a_J^h \geq b$. For every $j = 1, 2, \ldots, J_h$, we divide the interval $[\,a_{j-1}^h, a_j^h\,]$ in $K \geq 1$ subintervals $[\,x_j^{k-1}, x_j^k\,]$, $1 \leq k \leq K$, where

$$x_j^k \;=\; a_{j-1}^h \,+\, k\,\frac{h}{K}.$$

We then have, as $K \to \infty$,

$$\sum_{j=1}^{J_h} \frac{1}{K} \sum_{k=1}^{K} |\, f(x_j^k + h) - f(x_j^k)\,| \;\overset{K\to\infty}{\longrightarrow}\; \frac{1}{h} \int_a^{a_{J_h}^h} \big|\, f(x+h) - f(x)\,\big|\, dx.$$

On the other hand,

$$\begin{aligned}
\frac{1}{h} \int_a^{a_{J_h}^h} \big|\, f(x+h) - f(x)\,\big|\, dx &= \sum_{j=1}^{J_h} \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \big|\, f(x_{j+1}^k) - f(x_j^k)\,\big| \\
&\geq \sum_{j=1}^{J_h} \lim_{K\to\infty} \left|\, \frac{1}{K} \sum_{k=1}^{K} f(x_{j+1}^k) \;-\; \frac{1}{K} \sum_{k=1}^{K} f(x_j^k) \,\right| \\
&= \sum_{j=1}^{J_h} \left|\, M_{j+1}^h - M_j^h \,\right|,
\end{aligned}$$

where, for each $j = 1, 2, \ldots, J_h, J_h + 1, \ldots$, we set M_j^h to be the average values

$$M_j^h \;\equiv\; \frac{1}{h} \int_{a_{j-1}^h}^{a_j^h} f(x)\, dx\ .$$

In particular, defining $\mathcal{M}_h[\,f\,] : [\,a, b + h_0 - h\,] \mapsto R$ by putting, for $j = 1, 2, \ldots, J_h$, $J_h + 1, \ldots$, the values

$$\mathcal{M}_h[\,f\,](x) \;=\; \begin{cases} M_j^h & \text{if} \quad a_{j-1}^h < x < a_j^h, \\[2ex] \dfrac{M_j^h + M_{j+1}^h}{2} & \text{if} \quad x = a_j^h \end{cases}$$

and $\mathcal{M}_h[\,f\,](a) = M_1^h$, we get

$$\frac{1}{h}\int_a^{a^h_{J_h}} \big|\, f(x+h) - f(x)\,\big|\, dx \;\;\geq\;\; \mathrm{TV}\,[\,\mathcal{M}_h[\,f\,]\,;\, a, a^h_{J_h}+]. \tag{3.5}$$

Thus, (3.4) will be obtained if we show that

$$\liminf_{h\to 0^+}\; \mathrm{TV}\,[\,\mathcal{M}_h[\,f\,]\,;\, a, a^h_{J_h}+] \;\;\geq\;\; \mathrm{TV}\,[\,f\,;\, a, b+], \tag{3.6}$$

and this can be done as follows. Given $\eta > 0$ arbitrary, let $a = x_0 < x_1 < \cdots < x_{N-1} < x_N = b$ be such that

$$\text{(i)}\qquad \sum_{j=1}^{N} \big|\, f(x_j) - f(x_{j-1})\,\big| \;\;\geq\;\; \mathrm{TV}\,[\,f\,; a, b\,] \;-\; \eta,$$

$$\text{(ii)}\qquad f(b^-) \;-\; \eta \;\;\leq\;\; f(x) \;\;\leq\;\; f(b^-) \;+\; \eta \qquad \forall\, x \,\in\, [\,x_{N-1}, b\,[\;\;.$$

Since f is normal on $[\,a, b\,]$, we may also assume that

$$\text{(iii)}\qquad f \;\text{ is continuous at }\; x_1, x_2, \ldots, x_{N-1}.$$

Let then $\delta > 0$ be such that

$$\text{(iv)}\qquad f(x_j) \;-\; \frac{\eta}{N} \;\;\leq\;\; f(x) \;\;\leq\;\; f(x_j) \;+\; \frac{\eta}{N} \qquad \text{for each }\; x \,\in\, [\,x_j - \delta,\; x_j + \delta\,]$$

for every $j = 1, 2, \ldots, N-1$, and

$$\text{(v)}\qquad f(a) \;-\; \frac{\eta}{N} \;\;\leq\;\; f(x) \;\;\leq\;\; f(a) \;+\; \frac{\eta}{N} \qquad \text{for every }\; x \,\in\,]\,a,\; a+\delta\,],$$

$$\text{(vi)}\qquad f(b) \;-\; \frac{\eta}{N} \;\;\leq\;\; f(x) \;\;\leq\;\; f(b) \;+\; \frac{\eta}{N} \qquad \text{for every }\; x \,\in\,]\,b,\; b+2\,\delta\,],$$

$$\text{(vii)}\qquad \delta \;\;\leq\;\; \frac{b - x_{N-1}}{2}.$$

In particular, we get

$$\text{(a)}\qquad \mathcal{M}_h[\,f\,](x_j) \,\in\, \Big[\, f(x_j) - \frac{\eta}{N}\,,\; f(x_j) + \frac{\eta}{N}\,\Big] \quad \text{for }\; j = 1, 2, \ldots, N-1,$$

$$\text{(b)}\qquad \mathcal{M}_h[\,f\,](a) \,\in\, \Big[\, f(a) - \frac{\eta}{N}\,,\; f(a) + \frac{\eta}{N}\,\Big],$$

$$\text{(c)}\qquad \mathcal{M}_h[\,f\,]\bigg(a^h_{J_h} + \frac{h}{2}\bigg) \,\in\, \Big[\, f(b) - \frac{\eta}{N}\,,\; f(b) + \frac{\eta}{N}\,\Big],$$

and, by (ii), (iv), and (vi),

$$\text{(d)}\qquad \mathcal{M}_h[\,f\,](x_{N-1}) \,\in\, \Big[\, f(b^-) - \eta - \frac{\eta}{N}\,, f(b^-) + \eta + \frac{\eta}{N}\Big],$$

so that, for each $j = 1, 2, \ldots, N-1$,

$$\left|\, \mathcal{M}_h[\,f\,](x_j) \,-\, \mathcal{M}_h[\,f\,](x_{j-1}) \,\right| \;\geq\; \left|\, f(x_j) \,-\, f(x_{j-1}) \,\right| - \frac{2\,\eta}{N}$$

and

$$\left|\mathcal{M}_h[\,f\,]\left(a^h_{J_h} + \frac{h}{2}\right) - \mathcal{M}_h[\,f\,](x_{N-1})\right| \;\geq\; \left|\, f(x_N) \,-\, f(x_{N-1}) \,\right| - \frac{2\,\eta}{N} - \eta.$$

Hence, we obtain

$$\begin{aligned} &\mathrm{TV}\,[\,\mathcal{M}_h[\,f\,]\,;\, a, a^h_{J_h}+] \\ &\geq \left|\mathcal{M}_h[\,f\,]\left(a^h_{J_h} + \frac{h}{2}\right) - \mathcal{M}_h[\,f\,](x_{N-1})\right| + \sum_{j=1}^{N-1} \left|\, \mathcal{M}_h[\,f\,](x_j) \,-\, \mathcal{M}_h[\,f\,](x_{j-1}) \,\right| \\ &\geq \sum_{j=1}^{N} \left|\, f(x_j) \,-\, f(x_{j-1}) \,\right| \,-\, 3\,\eta\;, \end{aligned}$$

so that

$$\mathrm{TV}\,[\,\mathcal{M}_h[\,f\,]\,;\, a, a^h_{J_h}+] \;\geq\; \mathrm{TV}\,[\,f\,; a, b\,] \,-\, 4\,\eta$$

in view of (i). Since $\eta > 0$ is arbitrary, we then obtain

$$\liminf_{h \to 0^+} \, \mathrm{TV}\,[\,\mathcal{M}_h[\,f\,]\,;\, a, a^h_{J_h}+] \;\geq\; \mathrm{TV}\,[\,f\,;\, a, b\,],$$

which concludes the derivation of (3.6). Recalling (3.5), this gives (3.4). Finally, using (1.3), the inequalities (3.2b), (3.2c) are proved in a similar way. □

Clearly, this result establishes (1.10) at once. An easy way to see this is as follows: Given $f \in \mathrm{BV}(a,b)$, we extend it to $[\,a, +\infty\,[$ by setting $f(x) = f(b^-)$ for all $x > b$, and we redefine f on $[\,a, b\,]$ if necessary so that it be normal everywhere. Writing the integral in (3.2a) as a sum of two integrals corresponding to the intervals $[\,a, b-h\,]$ and $[\,b-h, b\,]$, we immediately get (1.10) from the limit (3.2a) and the fact that

$$\frac{1}{h} \int_{b-h}^{b} \left|\, f(x+h) - f(x) \,\right| dx \;\to\; 0 \qquad \text{as} \quad h \;\to\; 0$$

due to the continuity of f at the point b. Doing the same with the integrals (3.2b) and (3.2c), we then obtain the following result.

THEOREM 3.2. *Let $f \in \mathrm{BV}(a,b)$ be normal on $[\,a, b\,]$. Then*

$$\lim_{h \to 0^+} \frac{1}{h} \int_a^{b-h} \left|\, f(x+h) \,-\, f(x) \,\right| dx \;=\; \mathrm{TV}\,[\,f\,;\, a, b\,], \tag{3.7a}$$

$$\lim_{h \to 0^+} \frac{1}{h} \int_a^{b-h} \big(\, f(x+h) \,-\, f(x) \,\big)_+ \, dx \;=\; \mathrm{PV}\,[\,f\,;\, a, b\,], \tag{3.7b}$$

$$\lim_{h \to 0^+} \frac{1}{h} \int_a^{b-h} \big(\, f(x+h) \,-\, f(x) \,\big)_- \, dx \;=\; \mathrm{NV}\,[\,f\,;\, a, b\,]. \tag{3.7c}$$

As stated in the introduction, there is a converse to the result above, which we formulate as the following well-known property; see, e.g., [2], [13], [23], [35].

Theorem 3.3. *Let $f \in L^1(a,b)$ be such that*

$$\limsup_{h \to 0^+} \frac{1}{h} \int_a^{b-h} \big|\, f(x+h) - f(x) \,\big| \, dx \;=\; C \;<\; \infty. \tag{3.8}$$

Then f is of essential bounded variation on $[\,a,b\,]$ and $\mathrm{TV}_{ess}\,[\,f\,;\,a,b\,] \,=\, C$.

Proof. For the sake of completeness, we will provide a quick derivation of this result along the lines of the proof given in [2]; see also [8] for an alternative argument.

For each integer $n \geq 1$, let $h = (b-a)/n$ and $a_j^h = a + jh$, $0 \leq j \leq n$. We then construct the step function $\mathcal{M}_h[\,f\,] \in \mathrm{BV}(a,b)$ as in the proof of Theorem 3.1, i.e., we set

$$\mathcal{M}_h[\,f\,](x) \;=\; M_j^h \;=\; \frac{1}{h} \int_{a_{j-1}^h}^{a_j^h} f(s)\,ds \qquad \text{if} \quad a_{j-1}^h \,<\, x \,<\, a_j^h$$

for $j = 1, 2, \ldots, n$, extending it to the nodal points $\{\, a_0^h,\, a_1^h,\, \ldots\, , a_n^h \,\}$ in any way so as to become normal everywhere on $[\,a,b\,]$. Then

$$\mathrm{TV}\,[\,\mathcal{M}_h[\,f\,]\,;\,a,b\,] \;=\; \sum_{j=1}^{n} \big|\, M_j^h - M_{j-1}^h \,\big| \;\leq\; \frac{1}{h} \int_a^b \big|\, f(x+h) - f(x) \,\big| \, dx$$

so that, by (3.8), we obtain

$$\limsup_{h \to 0^+} \; \mathrm{TV}\,[\,\mathcal{M}_h[\,f\,]\,;\,a,b\,] \;\leq\; C, \tag{3.9}$$

from which we get that the total variation on $[\,a,b\,]$ of the family of functions $\mathcal{M}_h[\,f\,]$ is uniformly bounded as $h \,\to\, 0^+$. Moreover, it is well known that

$$\|\, \mathcal{M}_h[\,f\,] \,-\, f \,\|_{L^1(a,b)} \;\to\; 0 \tag{3.10}$$

as $h \,\to\, 0^+$; see, e.g., [2, Lemma 1.2.2]. This implies that some subsequence of the functions $\mathcal{M}_h[\,f\,]$ must converge pointwise to f a.e. on $[\,a,b\,]$ as $h \,\to\, 0^+$, say the sequence $\mathcal{M}_{h'}[\,f\,]$, and, because of (3.9), these functions $\mathcal{M}_{h'}[\,f\,]$ are then uniformly bounded on $[\,a,b\,]$. Thus, we can apply Helly's principle [22] and select a subsequence of $\mathcal{M}_{h'}[\,f\,]$, say $\mathcal{M}_{h''}[\,f\,]$, which converges pointwise to some function $g \in \mathrm{TV}(a,b)$ everywhere on $[\,a,b\,]$ as $h \to 0^+$. From (3.9), we immediately obtain that $\mathrm{TV}\,[\,g\,;\,a,b\,] \leq C$, while we also have $f = g$ a.e. on $[\,a,b\,]$ because of (3.10)—in fact, f is the a.e. pointwise limit of the whole sequence $\mathcal{M}_{h'}[\,f\,]$ as $h \to 0^+$. This is still true if we redefine g on $[\,a,b\,]$ so that it be normal everywhere on this interval. Now, from Theorem 3.2 and (3.8), we then have

$$\mathrm{TV}\,[\,g\,;\,a,b\,] \;=\; \lim_{h \to 0^+} \; \frac{1}{h} \int_a^{b-h} \big|\, g(x+h) - g(x) \,\big| \, dx \;=\; C,$$

since $g = f$ a.e. on $[\,a,b\,]$. This gives $\mathrm{TV}_{\mathrm{ess}}\,[\,f\,;\,a,b\,] \,=\, C$, as stated. $\square$

In fact, the proof given above establishes more, if we only recall Theorem 3.2. Given $f \in L^1(a,b)$ satisfying the condition (3.8), it follows that

$$\text{(3.11a)} \qquad \mathrm{TV}_{\mathrm{ess}}\,[\,f\,;\,a,b\,] \;=\; \lim_{h \to 0^+} \frac{1}{h} \int_a^{b-h} \big|\, f(x+h) - f(x) \,\big| \, dx,$$

$$\text{(3.11b)} \qquad \mathrm{PV}_{\mathrm{ess}}\,[\,f\,;\,a,b\,] \;=\; \lim_{h \to 0^+} \frac{1}{h} \int_a^{b-h} \big(\, f(x+h) - f(x) \,\big)_+ \, dx,$$

$$\text{(3.11c)} \qquad \mathrm{NV}_{\mathrm{ess}}\,[\,f\,;\,a,b\,] \;=\; \lim_{h \to 0^+} \frac{1}{h} \int_a^{b-h} \big(\, f(x+h) - f(x) \,\big)_- \, dx.$$

A note of caution should be given here. Given $f \in L^1(a, b + h_0)$ for some $h_0 > 0$, we observe that the property

$$\text{(3.12)} \qquad \limsup_{h \to 0^+} \frac{1}{h} \int_a^{b} \big|\, f(x+h) - f(x) \,\big| \, dx \;=\; C \;<\; \infty$$

does *not* imply that $\mathrm{TV}_{\mathrm{ess}}\,[\,f\,;\,a, b+\,] \,=\, C$ or the existence of the limit

$$\text{(3.13)} \qquad \lim_{h \to 0^+} \frac{1}{h} \int_a^{b} \big|\, f(x+h) - f(x) \,\big| \, dx \;=\; C.$$

We can ascertain from (3.12) only that f is of essential bounded variation on $[\,a,b\,]$ and that (3.11a), (3.11c) hold. In fact, the following example will illustrate this. Taking $0 < \varepsilon < 1$ and setting

$$\text{(3.14a)} \qquad f(x) \;=\; \begin{cases} 1 & \text{if} \quad \varepsilon^n \dfrac{1+\varepsilon}{1-\varepsilon} \;\le\; x \;\le\; \varepsilon^n \dfrac{2}{1-\varepsilon} \quad \text{for some } n \,\ge\, 1, \\[2ex] 0 & \text{otherwise,} \end{cases}$$

we obtain $f \in L^1(-1,1) \cap \mathrm{BV}(-1,0)$ with $\mathrm{TV}_{\mathrm{ess}}\,[\,f\,;\,-1\,,\,0+\,] = \infty$, and

$$\text{(3.14b)} \qquad \limsup_{h \to 0^+} \frac{1}{h} \int_{-1}^{0} \big|\, f(x+h) - f(x) \,\big| \, dx \;=\; \frac{1}{2}$$

while

$$\text{(3.14c)} \qquad \liminf_{h \to 0^+} \frac{1}{h} \int_{-1}^{0} \big|\, f(x+h) - f(x) \,\big| \, dx \;=\; \frac{\varepsilon}{1+\varepsilon} \;<\; \frac{1}{2}.$$

We conclude this discussion with one final remark. The results given above have been stated in terms of the differences $f(x+h) - f(x)$, but of course could have been written in terms of $f(x) - f(x-h)$ instead, with only some minor obvious changes in the statements or the analysis. Thus, for example, Theorem 3.1 could have been given as follows: given $f \in \mathrm{BV}(a - h_0, b)$ for some $h_0 > 0$,

$$\text{(3.15)} \qquad \lim_{h \to 0^+} \frac{1}{h} \int_a^{b} \big|\, f(x) - f(x-h) \,\big| \, dx \;=\; \mathrm{TV}_{\mathrm{ess}}\,[\,f\,;\,a-\,,b\,],$$

where

$$\mathrm{TV}_{\mathrm{ess}}\,[\,f\,;\,a-\,,b\,] \;=\; \lim_{\varepsilon \to 0^+} \mathrm{TV}_{\mathrm{ess}}\,[\,f\,;\,a-\varepsilon\,,b\,]$$

and similarly for the other left-side limits $\mathrm{PV}_{\mathrm{ess}}\,[\,f\,;\,a-\,,b\,]$ and $\mathrm{NV}_{\mathrm{ess}}\,[\,f\,;\,a-\,,b\,]$.

REFERENCES

[1] V. V. Chistyakov, *On mappings of bounded variation*, J. Dynam. Control Systems, 3 (1997), pp. 261–289.

[2] Z. Cybertowicz and W. Matuszewska, *Functions of bounded generalized variations*, Comment. Math. Prace Mat., 20 (1977), pp. 29–52.

[3] G. M. Ewing, *Calculus of Variations with Applications*, Norton, New York, 1969.

[4] H. Federer, *An analytic characterization of distributions whose partial derivatives are representable by measures*, Bull. Amer. Math. Soc., 60 (1954), p. 339.

[5] H. Federer, *Geometric Measure Theory*, Springer, New York, 1969.

[6] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, Wiley-Interscience, New York, 1984.

[7] A. M. Garsia and S. Sawyer, *On some classes of continuous functions with convergent Fourier series*, J. Math. Mech., 13 (1964), pp. 586–601.

[8] C. George, *Exercises in Integration*, Springer, New York, 1984.

[9] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser, Boston, 1984.

[10] E. Godlewski and P. A. Raviart, *Hyperbolic Systems of Conservation Laws*, Ellipses, Paris, France, 1991.

[11] E. Godlewski and P. A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Springer, New York, 1996.

[12] C. Goffman, *A geometric characterization of bounded variation and absolute continuity*, Bull. Inst. Math. Acad. Sinica, 9 (1981), pp. 395–398.

[13] G. H. Hardy and J. E. Littlewood, *Some properties of fractional integrals* I, Math. Z., 27 (1928), pp. 565–606.

[14] E. W. Hobson, *The Theory of Functions of a Real Variable and the Theory of Fourier Series*, Vol. I, 3rd ed., Cambridge University Press, Cambridge, UK, 1927.

[15] C. Jordan, *Sur la série de Fourier*, C. R. Acad. Sci., 92 (1881), pp. 228–230.

[16] C. Jordan, *Cours d'Analyse*, Vol. I, 2nd ed., Gauthier-Villars, Paris, France, 1893.

[17] R. Kannan and C. K. Krueger, *Advanced Analysis on the Real Line*, Springer, New York, 1996.

[18] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, Wiley, New York, 1967.

[19] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser, Boston, 1990.

[20] J. Málek, J. Nečas, M. Rokyta, and M. Růžička, *Weak and Measure-Valued Solutions to Evolutionary PDEs*, Chapman & Hall, London, UK, 1996.

[21] J. Musielak and W. Orlicz, *On generalized variations* I, Studia Math., 18 (1959), pp. 11–41.

[22] I. P. Natanson, *Theory of Functions of a Real Variable*, Ungar, New York, 1961.

[23] W. Orlicz, *On functions of finite variation depending on a parameter*, Studia Math., 13 (1953), pp. 218–232.

[24] M. Picone, *Sulla variazione totale di una funzione metrica*, Rend. Sem. Mat. Fis. Milano, 30 (1960), pp. 59–92.

[25] L. Schwartz, *Théorie des Distribuitions*, Hermann, Paris, France, 1966.

[26] L. Schwartz, *Analyse Mathématique*, Vol. I, Hermann, Paris, France, 1967.

[27] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, 2nd ed., Springer, New York, 1994.

[28] A. I. Vol'pert, *The spaces BV and quasilinear equations*, Mat. Sb. (N.S.), 2 (1967), pp. 225–267.

[29] A. I. Vol'pert and S. I. Hudjaev, *Cauchy's problem for degenerate second order quasilinear parabolic equations*, Math. USSR Sb., 7 (1970), pp. 365–387.

[30] D. Waterman, *On convergence of Fourier series of functions of generalized bounded variation*, Studia Math., 44 (1972), pp. 107–117 and p. 651 (errata).

[31] N. Wiener, *The quadratic variation of a function and its Fourier coefficients*, J. Mass. Inst. Tech., 3 (1924), pp. 73–94.

[32] L. C. Young, *Sur une généralisation de la notion de puissance p-ième bornée au sense de M. Wiener, et sur la convergence des séries de Fourier*, C. R. Acad. Sci. Paris, 204 (1937), pp. 470–472.

[33] W. H. Young, *On multiple Fourier series*, Proc. London Math. Soc. (2), 11 (1913), pp. 133–184.

[34] W. P. Ziemer, *Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation*, Springer, New York, 1989.

[35] A. Zygmund, *Trigonometric Series*, Vol. I, Cambridge University Press, Cambridge, UK, 1958.

SIAM J. MATH. ANAL.
Vol. 33, No. 5, pp. 1211–1227

ON THE STOCHASTIC EULER EQUATIONS IN A TWO-DIMENSIONAL DOMAIN[*]

JONG UHN KIM[†]

Abstract. In this paper, we discuss an initial-boundary value problem associated with the Euler equations with a random noise in a simply connected two-dimensional bounded domain. We present two different results according to the space regularity of the random noise. When the random noise is regular in the space variables, we prove the existence of a unique solution as a Banach space-valued continuous stochastic process. If the noise is less regular in the space variables, we establish the existence of solutions defined over a given probability space under the assumption that the noise is given in terms of a standard Brownian motion.

Key words. Euler equations, vorticity, existence, pathwise solutions, measurability, Brownian motion

AMS subject classifications. 35Q35, 76B03, 60H15

PII. S0036141001383941

Introduction. In this paper we will discuss an initial-boundary value problem for the Euler equations with a random noise in a two-dimensional simply connected domain. The problem is formulated as follows:

$$\frac{\partial u}{\partial t} + u \cdot \nabla u + \nabla p = \frac{\partial W}{\partial t} \qquad \text{for } (t,x) \in (0,T) \times G, \tag{0.1}$$

$$\nabla \cdot u = 0 \qquad \text{for } (t,x) \in (0,T) \times G, \tag{0.2}$$

$$u \cdot n = 0 \qquad \text{for } (t,x) \in [0,T] \times \partial G, \tag{0.3}$$

$$u(0,x) = u_0(x) \qquad \text{for } x \in G, \tag{0.4}$$

where $u = (u_1, u_2)$ is the velocity vector, p is the pressure, G is a simply connected bounded domain in R^2 with smooth boundary ∂G, and $u \cdot n$ stands for the normal component of u on ∂G. The right-hand side of (0.1) represents an external random noise. A Wiener process with values in a Hilbert space is a typical example for W.

The deterministic Euler equations have been extensively investigated. For a complete survey of known results for the deterministic Euler equations, see Lions [8] and references therein. We will recall only the results on the existence of solutions that are directly relevant to our investigation. The fundamental result on the existence and uniqueness of solutions for a two-dimensional domain is due to Yudovich [12]. The approach in [12] is to start with the vorticity equation in terms of the stream function. By the method of parabolic regularization, the existence of vorticity was established with the L^∞-estimate. This leads to the existence and uniqueness of the velocity vector. The case of a simply connected domain was analyzed first, and the analysis was extended to a multiply connected domain by introducing some auxiliary functions. Bardos [1] employed a different approach for a similar result. The method of [1] is to approximate the Euler equations directly by the Navier–Stokes equations

[*] Received by the editors January 19, 2001; accepted for publication (in revised form) November 6, 2001; published electronically April 8, 2002.
http://www.siam.org/journals/sima/33-5/38394.html

[†] Department of Mathematics, Virginia Polytechnic Institute and State University, 460 McBryde Hall, Blacksburg, VA 24061 (kim@math.vt.edu).

with special boundary conditions. The solution of the Euler equations was obtained as the zero viscosity limit. The method does not require geometric conditions on the space domain other than some smoothness of the boundary.

For the stochastic Euler equations, there are rather a limited number of results that include Bessaih [3], Bessaih and Flandoli [4], Brzezniak and Peszat [5], Capinski and Cutland [6], and Mikulevicius and Valiukevicius [9]. These works present results on the existence of solutions, which are different from each other. In particular, the issues raised in [4] have motivated our work. They [4] used the same approximation scheme as in [1] for the proof of existence. In fact, they could have relaxed regularity conditions on W in the space variables for the uniqueness of solutions if the viscosity term had been modified to contain W. But the boundary conditions are still required. This is due to the introduction of the viscosity term in (0.1). Here we propose to use a different method in the case of a simply connected domain. We will mainly work with the vorticity equation as in [12], but our approximation scheme is different. By virtue of this new scheme, we cannot only relax conditions on the regularity of $W(t,x)$ in the space variables x but also dispense with any boundary condition on W for the existence and uniqueness of solutions. One could imitate the procedure in [12] with some modification for the existence of solutions. But then the measurability of solutions as random variables is difficult to come by because a fixed point theorem was used. For a multiply connected domain, we can modify the system of approximate equations to include several unknown scalar functions corresponding to the inner boundaries following the idea of [12], and proceed in the same way as in the case of a simply connected domain. But the analysis will be far more lengthy, and we will not pursue this.

We will also discuss the case where W is less regular in the space variables and the solution is not known to be unique. This is the case where martingale solutions were obtained in [3]. Our goal is to find solutions over a given probability space with less regular W. We will show that this is possible if W is given in terms of a standard Brownian motion. When W is less regular in the space variables, we can still obtain pathwise solutions. But the uniqueness of solution is not known, and this causes difficulty in the measurability of pathwise solutions. For the stochastic Navier–Stokes equations, Bensoussan and Temam [2] overcame this difficulty by making a special assumption on the probability space and using a selection theorem. Here we use a general probability space, but W has to be expressed in terms of a Brownian motion. We will first work on the completion of the canonical probability space, from which the general case follows. Our crucial tool is a selection theorem proved in [2]. However, our approach is quite different because of insufficient regularity of W in the space variables. In fact, our procedure can relax the regularity condition on the example (3.38) in [2].

Our results are stated in section 1 and proved in sections 2 and 3.

1. Notation and statement of the results.

Throughout this paper, $T > 0$ is a given positive number, G is a simply connected bounded domain in R^2 with smooth boundary ∂G, and $H^\alpha(G)$ denotes the usual Sobolev space of order $\alpha \in R$. We will use the same notation for both vector-valued function classes and scalar-valued function classes. The function spaces $\mathcal{V}$ and $\mathcal{H}$ are defined by

$$\mathcal{V} = \big\{ f = (f_1, f_2) \,\big|\, f \in H^1(G),\ \nabla \cdot f = 0 \text{ in } G, \text{and } f \cdot n = 0 \text{ on } \partial G \big\},$$

$$\mathcal{H} = \big\{ f = (f_1, f_2) \,\big|\, f \in L^2(G),\ \nabla \cdot f = 0 \text{ in } G, \text{and } f \cdot n = 0 \text{ on } \partial G \big\}.$$

We will also use the notation $\nabla \times f = \partial f_2/\partial x_1 - \partial f_1/\partial x_2$.

$(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ is a given stochastic basis, where P is a probability measure, $\mathcal{F}$ is a σ-algebra, and $\{\mathcal{F}_t\}_{t\geq 0}$ is a right-continuous filtration on $(\Omega, \mathcal{F})$ such that $\mathcal{F}_0$ contains all P-negligible subsets. When $\mathcal{X}$ is a metric space, the mapping $f : \Omega \to \mathcal{X}$ is said to be $\mathcal{X}$-valued $\mathcal{F}$-measurable if $f^{-1}(\mathcal{O}) \in \mathcal{F}$ for every open subset $\mathcal{O} \subset \mathcal{X}$.

We will present two results under different assumptions.

1.1. The first result. We suppose that the random vector function $W = (W_1(t,x;\omega), W_2(t,x;\omega))$ satisfies the following conditions:

$$\text{For } P\text{-almost all } \omega \in \Omega,\ W(\cdot,\cdot\,;\omega) \in C([0,T]; H^{3+\alpha}(G)) \text{ for some } \alpha > 0. \tag{1.1}$$

$$\text{For each } t \in [0,T],\ W(t,\cdot\,;\cdot) \text{ is } H^{3+\alpha}(G)\text{-valued } \mathcal{F}_t\text{-measurable.} \tag{1.2}$$

Definition 1.1. *A random function $u(t,x;\omega)$ is said to be a solution of* (0.1)–(0.4) *if u is a $\mathcal{V}$-valued continuous stochastic process adapted to $\{\mathcal{F}_t\}_{0\leq t\leq T}$, and for P-almost all ω,* (0.1) *is satisfied with some distribution p in the sense of distribution over $(0,T)\times G$ and* (0.4) *holds.*

Under the above assumptions, our result is the following.

Theorem 1.2. *Suppose u_0 is $\mathcal{V}$-valued $\mathcal{F}_0$-measurable such that $\nabla \times u_0(\cdot\,;\omega) \in L^\infty(G)$ for P-almost all ω. Then, there is a unique solution to* (0.1)–(0.4).

Here uniqueness means pathwise uniqueness, and Definition 1.1 is used only for Theorem 1.2.

1.2. The second result.

1.2.1. A general probability space. We assume that W is given by

$$W(t,x;\omega) = g(x)B_t(\omega), \tag{1.3}$$

where g is R^2-valued and belongs to $H^1(G)$, and $B_t(\cdot)$ is a standard one-dimensional Brownian motion over $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$.

Theorem 1.3. *Suppose u_0 is deterministic and belongs to $\mathcal{V}$. Let $\delta \in (1/2, 0)$. Then, there is a random function u defined over Ω such that u is $L^2(0,T;\mathcal{V}) \cap C([0,T]; \mathcal{H}\cap H^\delta(G))$-valued $\mathcal{F}$-measurable and such that* (0.1)–(0.4) *are satisfied with some distribution p for P-almost all ω.*

This result will be first proved in the following special case.

1.2.2. The canonical probability space. Here we choose $\Omega = C([0,\infty))$, which is a Polish space under a standard metric. Let $\mathcal{B}(\Omega)$ be the σ-algebra of all Borel subsets of $C([0,\infty))$. Let P be the Wiener measure on $(\Omega, \mathcal{B}(\Omega))$ such that the coordinate mapping process

$$X_t : \omega \mapsto \omega(t) \tag{1.4}$$

is the standard Brownian motion under P. Then, $\big(\Omega,\, \mathcal{B}(\Omega), P\big)$ is called the canonical probability space. Following Karatzas and Shreve [7], we define

$$\mathcal{N} = \left\{ A \subset \Omega \,\middle|\, \text{there is a } B \in \mathcal{B}(\Omega) \text{ with } A \subset B \text{ and } P(B) = 0 \right\}, \tag{1.5}$$

$$\mathcal{F}_t = \text{the } \sigma\text{-algebra generated by } X_s,\ 0 \leq s \leq t, \text{ and } \mathcal{N}, \tag{1.6}$$

$$\mathcal{F} = \text{the completion of } \mathcal{B}(\Omega) \text{ under } P. \tag{1.7}$$

Then, $\{\mathcal{F}_t\}$ is a right-continuous filtration on $(\Omega, \mathcal{F})$ (see [7, p. 90]), and $\mathcal{F}_0$ contains all P-negligible subsets. Furthermore, X_t is again a Brownian motion on $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$. We assume that

$$W(t,x;\omega) = g(x)X_t(\omega), \tag{1.8}$$

where g is R^2-valued and belongs to $H^1(G)$.

2. Proof of Theorem 1.2. We first choose a $\mathcal{F}_0$-measurable set $\tilde{\Omega} \subset \Omega$ with $P(\tilde{\Omega}) = 1$ such that for each $\omega \in \tilde{\Omega}$, $W(\cdot,\, \cdot;\omega) \in C([0,T]; H^{3+\alpha}(G))$, $\nabla \times u_0(\cdot\,;\omega) \in L^\infty(G)$, and $u_0(\cdot\,;\omega) \in \mathcal{V}$ hold. We consider the following initial-boundary value problem for a scalar random function $\zeta_\epsilon(t,x;\omega)$ with parameter $\epsilon > 0$:

$$\frac{\partial \zeta_\epsilon}{\partial t} + (\nabla^\perp \Psi_\epsilon)\cdot\nabla\zeta_\epsilon = -(\nabla^\perp\Psi_\epsilon)\cdot\nabla(\nabla\times W_\epsilon(t,x;\omega)) \quad \text{for } (t,x) \in (0,T)\times G, \tag{2.1}$$

$$\zeta_\epsilon(0,x;\omega) = \nabla\times u_{0,\epsilon}(x;\omega) - \nabla\times W_\epsilon(0,x;\omega) \quad \text{for } x \in G, \tag{2.2}$$

where $\nabla^\perp = (\partial/\partial x_2, -\partial/\partial x_1)$, and Ψ_ϵ is the solution of

$$\begin{cases} -(1-\epsilon\Delta)\Delta\Psi_\epsilon = \zeta_\epsilon + \nabla\times W_\epsilon(t,\cdot;\omega) \quad \text{in } G, \\ \Psi_\epsilon = \Delta\Psi_\epsilon = 0, \quad \text{on } \partial G. \end{cases} \tag{2.3}$$

Equation (2.1) is a regularized vorticity equation. Here W_ϵ is defined by

$$W_\epsilon = r_G\left((\mathcal{L}W)\star\rho_\epsilon\right), \tag{2.4}$$

where r_G denotes restriction to G, $\mathcal{L}$ is a continuous extension: $H^{3+\alpha}(G) \to H^{3+\alpha}(R^2)$ such that $r_G\,\mathcal{L}h = h$ for all $h \in H^{3+\alpha}(G)$, and ρ_ϵ is the Friedrichs mollifier. For $\mathcal{L}$, it is sufficient that ∂G is C^m-smooth with $m \geq 3+\alpha$. The convolution is taken with respect to the space variables $x \in R^2$. Then, for each $\omega \in \tilde{\Omega}$, $W_\epsilon \in C([0,T]; C^3(\overline{G}))$ and

$$W_\epsilon \to W \text{ strongly in } C([0,T]; H^{3+\alpha}(G)) \text{ as } \epsilon \to 0. \tag{2.5}$$

In the meantime, $u_{0,\epsilon}$ in (2.2) is an approximation of u_0 such that $u_{0,\epsilon} \in C^2(\overline{G}) \cap \mathcal{V}$ and, for each $\omega \in \tilde{\Omega}$, as $\epsilon \to 0$,

$$\nabla\times u_{0,\epsilon} \;\to\; \nabla\times u_0 \qquad \begin{cases} \text{weak-star in } L^\infty(G), \\ \text{strongly in } L^2(G). \end{cases} \tag{2.6}$$

We will sketch a procedure to construct such approximations. Let us define

$$h(x;\omega) = \begin{cases} \nabla\times u_0(x;\omega) \quad \text{for } x \in G, \\ 0 \quad \text{for } x \notin G \end{cases} \tag{2.7}$$

and

$$h_\epsilon = \rho_\epsilon \star h. \tag{2.8}$$

Then $h_\epsilon \in C^\infty(R^2)$, and as $\epsilon \to 0$, $h_\epsilon \to h$ weak-star in $L^\infty(G)$ and strongly in $L^2(R^2)$ for each $\omega \in \tilde{\Omega}$. We then find Φ_ϵ satisfying

$$\begin{cases} -\Delta\Phi_\epsilon = h_\epsilon \quad \text{in } G, \\ \Phi_\epsilon = 0 \quad \text{on } \partial G. \end{cases} \tag{2.9}$$

We now choose $u_{0,\epsilon} = \nabla^{\perp}\Phi_{\epsilon}$. Since $h_{\epsilon} \in C^{\infty}(\overline{G})$, it follows that $\Phi_{\epsilon} \in C^{\infty}(\overline{G})$. We also note that $\nabla \times (\nabla^{\perp}\Phi_{\epsilon}) = -\Delta\Phi_{\epsilon} = h_{\epsilon}$ in G and that $n \cdot \nabla^{\perp}\Phi_{\epsilon} = \partial\Phi_{\epsilon}/\partial s = 0$ on ∂G. Here n is a normal vector and $\partial/\partial s$ is tangential differentiation on ∂G. Hence, $u_{0,\epsilon}$ satisfies all the required properties.

LEMMA 2.1. *For each $\omega \in \tilde{\Omega}$, there is a unique solution $\zeta_{\epsilon}(\cdot\,,\cdot\,;\omega) \in C^1([0,T] \times \overline{G})$ of (2.1)–(2.3). Furthermore, for each $0 < s \leq T$, $\zeta_{\epsilon}(s,\cdot)$ is $L^2(G)$-valued $\mathcal{F}_s$-measurable, and it holds that*

$$\|\zeta_{\epsilon}\|_{C([0,T]\times\overline{G})} \leq M(\omega), \tag{2.10}$$

$$\left\|\frac{\partial\zeta_{\epsilon}}{\partial t}\right\|_{C([0,T];H^{-1}(G))} \leq M(\omega) \tag{2.11}$$

for some positive constants $M(\omega)$ independent of ϵ.

Proof. For approximate solutions, we use a complete orthonormal system of the eigenfunctions $\{e_m\}_{m=1}^{\infty}$ of

$$\begin{cases} -\Delta e_m = \lambda_m e_m & \text{in } G, \\ e_m = 0 \quad \text{on } \partial G. \end{cases} \tag{2.12}$$

Let us write

$$\zeta_{\epsilon,m}(t,x;\omega) = \sum_{k=1}^{m} c_{m,k}(t;\omega)e_k(x) \tag{2.13}$$

and define $\Psi_{\epsilon,m}$ to be the solution of

$$\begin{cases} -(1-\epsilon\Delta)\Delta\Psi_{\epsilon,m} = \zeta_{\epsilon,m} + \displaystyle\sum_{k=1}^{m} \langle\nabla \times W_{\epsilon}(t,\cdot\,;\omega),\, e_k\rangle_{L^2(G)}\, e_k & \text{in } G, \\ \Psi_{\epsilon,m} = \Delta\Psi_{\epsilon,m} = 0, \quad \text{on } \partial G, \end{cases} \tag{2.14}$$

where $\langle\cdot\,,\,\cdot\rangle_{L^2(G)}$ is the inner product in $L^2(G)$. We then consider the following system of ordinary differential equations:

$$\begin{aligned} \frac{dc_{m,k}}{dt} &+ \langle(\nabla^{\perp}\Psi_{\epsilon,m}) \cdot \nabla\zeta_{\epsilon,m},\, e_k\rangle_{L^2(G)} \\ &= \langle-(\nabla^{\perp}\Psi_{\epsilon,m}) \cdot \nabla(\nabla \times W_{\epsilon}(t,\cdot\,;\omega)),\, e_k\rangle_{L^2(G)}, \quad k = 1,\ldots,m, \end{aligned} \tag{2.15}$$

$$c_{m,k}(0;\omega) = \langle\nabla \times u_{0,\epsilon}(\cdot\,;\omega) - \nabla \times W_{\epsilon}(0,\cdot\,;\omega),\, e_k\rangle_{L^2(G)}, \quad k = 1,\ldots,m. \tag{2.16}$$

The system (2.15) can be put in the form

$$\frac{dz}{dt} = h(t,z;\omega), \tag{2.17}$$

where $z = (c_{m,1},\ldots,c_{m,m})$, and all the components of the vector function h are polynomials in $c_{m,k}$'s with coefficients depending on t, ω. For each $\omega \in \tilde{\Omega}$, the coefficients belong to $C([0,T])$. Thus, the existence and uniqueness of local solutions follow immediately. For the existence on the whole interval $[0,T]$, we need some a priori estimates. We first note that, for each t,

$$\langle(\nabla^{\perp}\Psi_{\epsilon,m}) \cdot \nabla\zeta_{\epsilon,m},\, \zeta_{\epsilon,m}\rangle_{L^2(G)} = 0 \tag{2.18}$$

and

$$\|\nabla^\perp \Psi_{\epsilon,m}\|_{L^2(G)} \le M\Big(\|\zeta_{\epsilon,m}\|_{L^2(G)} + \|\nabla \times W_\epsilon\|_{L^2(G)} \Big), \tag{2.19}$$

which implies

$$\begin{aligned} &\Big| \langle -(\nabla^\perp \Psi_{\epsilon,m}) \cdot \nabla(\nabla \times W_\epsilon),\, \zeta_{\epsilon,m}\rangle_{L^2(G)} \Big| \\ &\quad \le M \|\nabla^\perp \Psi_{\epsilon,m}\|_{L^2(G)} \|\zeta_{\epsilon,m}\|_{L^2(G)} \|\nabla(\nabla \times W_\epsilon)\|_{L^\infty(G)} \\ &\quad \le M\Big(\|\zeta_{\epsilon,m}\|^2_{L^2(G)} + \|\nabla \times W_\epsilon\|^2_{L^2(G)} \Big) \|W_\epsilon\|_{H^{3+\alpha}(G)}, \end{aligned} \tag{2.20}$$

where M denotes positive constants independent of t, m, and ϵ. We multiply (2.15) by $c_{m,k}$, sum over $1 \le k \le m$, and apply Gronwall's inequality to obtain bounds for $c_{m,k}$'s. It follows that for each $\omega \in \tilde{\Omega}$, the above system has a unique solution in $C^1([0,T])$, and

$$\sup_{0\le t\le T} \|\zeta_{\epsilon,m}(t)\|_{L^2(G)} \le M(\omega), \tag{2.21}$$

where $M(\omega)$ is a positive constant independent of ϵ and m. Later, we will need the estimates of the time derivative of $\zeta_{\epsilon,m}$. For this, we note that

$$\begin{aligned} &\Big\| \sum_{k=1}^m \langle (\nabla^\perp \Psi_{\epsilon,m}) \cdot \nabla \zeta_{\epsilon,m},\, e_k \rangle_{L^2(G)}\, e_k \Big\|_{L^\infty(0,T;H^{-1}(G))} \\ &\le K \|(\nabla^\perp \Psi_{\epsilon,m}) \cdot \nabla \zeta_{\epsilon,m}\|_{L^\infty(0,T;H^{-1}(G))} = K \|\nabla \cdot \big((\nabla^\perp \Psi_{\epsilon,m})\zeta_{\epsilon,m}\big)\|_{L^\infty(0,T;H^{-1}(G))} \\ &\le K \|(\nabla^\perp \Psi_{\epsilon,m})\zeta_{\epsilon,m}\|_{L^\infty(0,T;L^2(G))} \le M(\epsilon,\omega), \end{aligned} \tag{2.22}$$

where K denotes a positive constant depending only on G, and $M(\epsilon, \omega)$ is a positive constant independent of m. Here we have used the fact that $H^{-1}(G)$ can be characterized by

$$H^{-1}(G) = \left\{ \sum_{k=1}^\infty a_k e_k \,\middle|\, \sum_{k=1}^\infty |a_k|^2/\lambda_k < \infty \right\}.$$

We also find that

$$\begin{aligned} &\Big\| \sum_{k=1}^m \langle (\nabla^\perp \Psi_{\epsilon,m}) \cdot \nabla(\nabla \times W_\epsilon),\, e_k \rangle_{L^2(G)}\, e_k \Big\|_{L^\infty(0,T;L^2(G))} \\ &\le K \|\nabla^\perp \Psi_{\epsilon,m}\|_{L^2(G)} \|W_\epsilon\|_{H^{3+\alpha}(G)} \le M(\omega), \end{aligned} \tag{2.23}$$

where K and $M(\omega)$ are positive constants independent of ϵ and m. It follows from (2.15) and the above estimates that

$$\left\| \frac{\partial \zeta_{\epsilon,m}}{\partial t} \right\|_{L^\infty(0,T;H^{-1}(G))} \le M(\epsilon,\omega). \tag{2.24}$$

For the measurability of $c_{m,k}$'s, we recall that the components of h in (2.17) are polynomials in $c_{m,k}$'s, where all the coefficients belong to $C([0,T])$ for each $\omega \in \tilde{\Omega}$, and depend continuously on W_ϵ in the sense that the mapping $W_\epsilon(\cdot,\cdot;\omega) \mapsto$ each coefficient is continuous from $C\big([0,T]; H^{3+\alpha}(G)\big)$ to $C([0,T])$ for each $\omega \in \tilde{\Omega}$. Since $W_\epsilon(s,\cdot;\cdot)$ is $H^{3+\alpha}(G)$-valued $\mathcal{F}_s$-measurable, it follows that $c_{m,k}(s;\cdot)$'s are $\mathcal{F}_s$-measurable for each $s \in (0,T]$. Now we proceed to construct pathwise solutions. Let us fix $\omega \in \tilde{\Omega}$. By virtue of (2.21) and (2.24), we can extract a subsequence still denoted by $\{\zeta_{\epsilon,m}\}_{m=1}^{\infty}$ such that as $m \to \infty$,

$$\zeta_{\epsilon,m} \to \zeta_\epsilon \quad \text{weak-star in } L^\infty(0,T;L^2(G)), \tag{2.25}$$

and

$$\zeta_{\epsilon,m} \to \zeta_\epsilon \quad \text{strongly in } C([0,T];H^{-\beta}(G)) \text{ for any } \beta \in (0,1/2) \tag{2.26}$$

for some $\zeta_\epsilon \in L^\infty(0,T;L^2(G))$ with $\partial\zeta_\epsilon/\partial t \in L^\infty(0,T;H^{-1}(G))$. For compactness in $C\big([0,T];H^{-\beta}(G)\big)$, see Corollary 8 of Simon [10]. This together with (2.14) yields

$$\Psi_{\epsilon,m} \to \Psi_\epsilon \quad \text{strongly in } C([0,T];H^{-\beta+4}(G)), \tag{2.27}$$

where Ψ_ϵ satisfies

$$\begin{cases} -(1-\epsilon\Delta)\Delta\Psi_\epsilon = \zeta_\epsilon + \nabla\times W_\epsilon(t,\cdot;\omega) \quad \text{in } G, \\ \Psi_\epsilon = \Delta\Psi_\epsilon = 0 \quad \text{on } \partial G. \end{cases} \tag{2.28}$$

It is easy to see that as $m \to \infty$,

$$(\nabla^\perp\Psi_{\epsilon,m})\cdot\nabla\zeta_{\epsilon,m} \to \nabla^\perp\Psi_\epsilon\cdot\nabla\zeta_\epsilon \quad \text{weak-star in } L^\infty(0,T;H^{-1}(G)). \tag{2.29}$$

Hence, ζ_ϵ is a solution of (2.1) and (2.2) for each $\omega \in \tilde{\Omega}$. We now show that this ζ_ϵ is necessarily in $C^1([0,T]\times\overline{G})$. For this, we consider the following linear deterministic problem:

$$\frac{\partial\zeta}{\partial t} + \Phi\cdot\nabla\zeta = F, \tag{2.30}$$

$$\zeta(0,\cdot) = \zeta_0. \tag{2.31}$$

Here Φ, F, and ζ_0 are given functions that satisfy the following conditions:

(i) Φ is R^2-valued and belongs to $C\big([0,T];C^1(\overline{G})\big)$.

(ii) Φ is tangential on the boundary ∂G for all t, and $\nabla\cdot\Phi = 0$ for all (t,x).

(iii) $F \in C\big([0,T];C^1(\overline{G})\big)$ and $\zeta_0 \in C^1(\overline{G})$.

Under these conditions, we have the following lemma.

Lemma 2.2. *Suppose that $\zeta \in L^\infty(0,T;L^2(G))$ is a solution of* (2.30)–(2.31). *Then $\zeta \in C^1([0,T]\times\overline{G})$.*

Proof. Since $\zeta \in L^\infty(0,T;L^2(G))$ satisfies (2.30) in the sense of distribution over $(0,T)\times G$, it holds that $\zeta \in C([0,T];H^{-\gamma}(G))$ for any $\gamma > 0$. Thus, (2.31) makes sense. By virtue of the conditions (i) and (ii), the flow associated with

$$\begin{cases} \dfrac{dx}{dt} = \Phi(t,x), \\ x(0) = y \in \overline{G} \end{cases} \tag{2.32}$$

exists on $[0,T]$ and gives rise to a volume preserving C^1-diffeomorphism $y \mapsto x = \eta(t,y)$ for each $t \in [0,T]$. Choose any arbitrary $\phi \in C_0^1((0,T)\times G)$. It follows from (2.30) and the divergence-free condition of Φ that

$$\int_0^T \int_G \zeta\Big(\frac{\partial \phi}{\partial t} + \Phi\cdot\nabla\phi\Big)\,dx\,dt = -\int_0^T\int_G F\phi\,dx\,dt. \tag{2.33}$$

We now express the integrals in terms of the new variables (s,y), where $t = s$ and $x = \eta(s,y)$:

$$\begin{aligned}\int_0^T \int_G \zeta(t,x)\Big(\frac{\partial \phi}{\partial t}(t,x) + \Phi(t,x)\cdot\nabla\phi(t,x)\Big)\,dx\,dt \\ = \int_0^T\int_G \zeta(s,\eta(s,y))\frac{\partial\phi}{\partial s}(s,\eta(s,y))\,dy\,ds,\end{aligned} \tag{2.34}$$

$$\int_0^T\int_G F(t,x)\phi(t,x)\,dx\,dt = \int_0^T\int_G F(s,\eta(s,y))\phi(s,\eta(s,y))\,dy\,ds. \tag{2.35}$$

Thus, it holds that

$$\frac{\partial\zeta}{\partial s}(s,\eta(s,y)) = F(s,\eta(s,y)) \tag{2.36}$$

in the sense of distribution over $(0,T)\times G$. We note that as a function of (s,y), $\zeta(s,\eta(s,y))$ belongs to $L^\infty\big(0,T;L^2(G)\big)$ because $y\mapsto\eta(s,y)$ is volume preserving for each s. Meanwhile,

$$\zeta(0,\eta(0,y)) = \zeta(0,y) = \zeta_0(y). \tag{2.37}$$

Since the initial value problem

$$\begin{cases}\dfrac{\partial z}{\partial s}(s,y) = F(s,\eta(s,y)), \quad (s,y)\in(0,T)\times G,\\ z(0,y) = \zeta_0(y), \quad y\in G,\end{cases}$$

has at most one solution in the space of distributions over $(0,T)\times G$, ζ is necessarily represented by

$$\zeta(t,x) = \zeta_0(y) + \int_0^t F(s,\eta(s,y))\,ds \tag{2.38}$$

for each $(t,x)\in(0,T)\times G$, where $x = \eta(t,y)$. Since the right-hand side belongs to $C^1([0,T]\times\overline{G})$, so does ζ. This completes the proof of the lemma. □

For each fixed $\omega\in\tilde\Omega$, we set $\Phi(t,x) = \nabla^\perp\Psi_\epsilon(t,x;\omega)$, $F(t,x) = -(\nabla^\perp\Psi_\epsilon)\cdot\nabla(\nabla\times W_\epsilon(t,x;\omega))$, and $\zeta_0(x) = \nabla\times u_{0,\epsilon}(x;\omega) - \nabla\times W_\epsilon(0,x;\omega)$. Then, the above conditions (i), (ii), and (iii) are satisfied, and $\zeta_\epsilon\in L^\infty(0,T;L^2(G))$ is a solution of (2.30)–(2.31). By Lemma 2.2, ζ_ϵ is necessarily in $C^1([0,T]\times\overline{G})$.

Next we will show that ζ_ϵ is a unique solution of (2.1)–(2.3). Fix $\epsilon>0$, and let ζ and $\tilde\zeta$ be two weak solutions in $C([0,T];L^2(G))$ of (2.1)–(2.3), and let Ψ and $\tilde\Psi$ be associated with ζ and $\tilde\zeta$, respectively, through (2.3). Then, both ζ and $\tilde\zeta$ belong to

$C^1([0,T]\times\overline{G})$ by the same argument as above. We denote the right-hand side of (2.1) by F and $\tilde{F}$ associated with Ψ and $\tilde{\Psi}$, respectively. We then have

$$\frac{\partial(\zeta-\tilde{\zeta})}{\partial t}+(\nabla^{\perp}\Psi)\cdot\nabla(\zeta-\tilde{\zeta})+\nabla^{\perp}(\Psi-\tilde{\Psi})\cdot\nabla\tilde{\zeta}=F-\tilde{F} \tag{2.39}$$

and

$$\|\nabla^{\perp}(\Psi-\tilde{\Psi})\|_{C(\overline{G})}\le M(\epsilon,\omega)\|\zeta-\tilde{\zeta}\|_{L^2(G)}\quad\text{for all } t. \tag{2.40}$$

By multiplying (2.39) by $\zeta-\tilde{\zeta}$ and integrating over G, we can easily find

$$\zeta\equiv\tilde{\zeta}, \tag{2.41}$$

which shows the uniqueness. Hence ζ_ϵ is determined independently of any subsequence $\{\zeta_{\epsilon,m}\}_{m=1}^{\infty}$. Consequently, the whole sequence $\{\zeta_{\epsilon,m}\}_{m=1}^{\infty}$ converges to ζ_ϵ for each $\omega\in\tilde{\Omega}$. Next we will establish the measurability of ζ_ϵ as a random function. Fix any $0<s\le T$. Each $\zeta_{\epsilon,m}(s,\cdot)$ is $H^{\mu}(G)$-valued $\mathcal{F}_s$-measurable for each $\mu\in R$, which follows from the measurability of $c_{m,k}$'s and the structure of $\zeta_{\epsilon,m}$. By virtue of (2.26), $\zeta_\epsilon(s,\cdot)$ is $H^{-\beta}(G)$-valued $\mathcal{F}_s$-measurable for $\beta\in(0,1/2)$. But each closed ball in $L^2(G)$ is a Borel subset of $H^{-\beta}(G)$, and thus $\zeta_\epsilon(s,\cdot)$ is $L^2(G)$-valued $\mathcal{F}_s$-measurable.

Next we will obtain a priori estimates independent of ϵ. By multiplying (2.1) by ζ_ϵ^3 and integrating over G, we find

$$\frac{d}{dt}\int_G\zeta_\epsilon^4\,dx=\frac{1}{4}\int_G F_\epsilon\,\zeta_\epsilon^3\,dx\quad\text{for all }\omega\in\tilde{\Omega}. \tag{2.42}$$

It follows from (2.3) that

$$\|\Delta\Psi_\epsilon\|_{L^4(G)}\le M\big(\|\zeta_\epsilon\|_{L^4(G)}+\|\nabla\times W_\epsilon\|_{L^4(G)}\big), \tag{2.43}$$

which yields

$$\|\nabla^{\perp}\Psi_\epsilon\|_{C(\overline{G})}\le M\big(\|\zeta_\epsilon\|_{L^4(G)}+\|\nabla\times W_\epsilon\|_{L^4(G)}\big), \tag{2.44}$$

where M denotes positive constants independent of t and ϵ. We can derive from (2.42) with help of (2.5) and (2.44):

$$\frac{d}{dt}\int_G\zeta_\epsilon^4\,dx\le M(\omega)\int_G\zeta_\epsilon^4\,dx+M(\omega), \tag{2.45}$$

which, together with (2.6), implies

$$\|\zeta_\epsilon\|_{C([0,T];L^4(G))}\le M(\omega), \tag{2.46}$$

where, here and below, $M(\omega)$ stands for positive constants independent of ϵ. It now follows from (2.5) and (2.44) that

$$\|F_\epsilon\|_{C([0,T]\times\overline{G})}\le M(\omega), \tag{2.47}$$

which, combined with (2.6) and (2.38), yields

$$\|\zeta_\epsilon\|_{C([0,T]\times\overline{G})}\le M(\omega). \tag{2.48}$$

We can also derive from (2.1) that

$$\left\| \frac{\partial \zeta_\epsilon}{\partial t} \right\|_{C([0,T];H^{-1}(G))} \leq M(\omega). \tag{2.49}$$

This ends the proof of Lemma 2.1. □

We now proceed to obtain a solution of

$$\frac{\partial \zeta}{\partial t} + (\nabla^\perp \Psi) \cdot \nabla \zeta = -(\nabla^\perp \Psi) \cdot \nabla(\nabla \times W(t,x;\omega)) \quad \text{for } (t,x) \in (0,T) \times G, \tag{2.50}$$

$$\zeta(0,x;\omega) = \nabla \times u_0(x;\omega) - \nabla \times W(0,x;\omega) \quad \text{for } x \in G, \tag{2.51}$$

where Ψ is the solution of

$$\begin{cases} -\Delta \Psi = \zeta + \nabla \times W(t,\cdot\,;\omega) & \text{in } G, \\ \Psi = 0 \quad \text{on } \partial G. \end{cases} \tag{2.52}$$

For the time being, we establish only the existence of a solution for each $\omega \in \tilde{\Omega}$.

Lemma 2.3. *For each* $\omega \in \tilde{\Omega}$, *there is a solution* ζ *of* (2.50)–(2.52) *such that* $\zeta \in L^\infty((0,T) \times G) \cap C([0,T]; H^{-\beta}(G))$ *for any* $\beta \in (0,1/2)$.

Proof. Let us fix $\omega \in \tilde{\Omega}$. By virtue of (2.48) and (2.49), we can use Corollary 8 of [10] to extract a subsequence still denoted by $\{\zeta_\epsilon\}$ and its companion $\{\Psi_\epsilon\}$ through (2.3) such that as $\epsilon \to 0$,

$$\zeta_\epsilon \to \zeta \quad \text{weak-star in } L^\infty(0,T;L^\infty(G)), \tag{2.53}$$

$$\zeta_\epsilon \to \zeta \quad \text{strongly in } C([0,T];H^{-\beta}(G)) \ \text{ for any } \beta \in (0,1/2), \tag{2.54}$$

and

$$\Psi_\epsilon \to \Psi \quad \text{strongly in } C([0,T];H^{2-\beta}(G)) \tag{2.55}$$

for some ζ and Ψ which satisfy

$$\begin{cases} -\Delta \Psi = \zeta + \nabla \times W & \text{in } G, \\ \Psi = 0 \quad \text{on } \partial G. \end{cases} \tag{2.56}$$

Since $\nabla^\perp \Psi_\epsilon$ is divergence-free, $(\nabla^\perp \Psi_\epsilon) \cdot \nabla \zeta_\epsilon = \nabla \cdot (\zeta_\epsilon \nabla^\perp \Psi_\epsilon)$ holds. Meanwhile, we use (2.53) and (2.55) to find that as $\epsilon \to 0$,

$$\nabla \cdot (\zeta_\epsilon \nabla^\perp \Psi_\epsilon) \to \nabla \cdot (\zeta \nabla^\perp \Psi) \quad \text{weak-star in } L^\infty(0,T;H^{-1}(G)). \tag{2.57}$$

It follows that ζ satisfies (2.50)–(2.52), and the proof of Lemma 2.3 is complete. □

For each $\omega \in \tilde{\Omega}$, we use the above Ψ to set

$$v = \nabla^\perp \Psi \tag{2.58}$$

so that

$$\nabla \times v = -\Delta \Psi = \zeta + \nabla \times W, \tag{2.59}$$

and (2.50) can be written as

$$\frac{\partial}{\partial t}(\nabla \times v) + \nabla \cdot \big((\nabla \times v)v\big) = \frac{\partial}{\partial t}\nabla \times W. \tag{2.60}$$

By straightforward differentiation, it is easy to see

$$\nabla \times \big((v \cdot \nabla)v\big) = \nabla \cdot \big((\nabla \times v)v\big), \tag{2.61}$$

and thus (2.60) reduces to

$$\nabla \times \left(\frac{\partial v}{\partial t} + (v \cdot \nabla)v - \frac{\partial W}{\partial t}\right) = 0. \tag{2.62}$$

Since G is a simply connected domain, there is a scalar distribution p over $(0,T) \times G$ such that

$$\frac{\partial v}{\partial t} + (v \cdot \nabla)v + \nabla p = \frac{\partial W}{\partial t}. \tag{2.63}$$

In the meantime, (2.51) and (2.59) yield

$$\nabla \times v(0,\cdot) = \nabla \times u_0(\cdot\,;\omega), \tag{2.64}$$

which implies

$$v(0,\cdot) = u_0(\cdot\,;\omega) + \nabla q \tag{2.65}$$

for some scalar distribution q over G. Since $v(0,\cdot) \in \mathcal{H}$ and $u_0(\cdot\,;\omega) \in \mathcal{V}$, it holds that $\nabla q \in \mathcal{H}$, which implies $\nabla q \equiv 0$. Hence, we have shown that for fixed $\omega \in \tilde{\Omega}$, v satisfies (0.1)–(0.4). It follows from (2.53), (2.55), (2.56), and (2.58) that for each $\omega \in \tilde{\Omega}$, $v \in C([0,T];\mathcal{H}) \cap L^\infty(0,T;\mathcal{V})$ and, by a result in [12],

$$\left\| \frac{\partial v}{\partial x_i} \right\|_{L^\infty(0,T;L^r(G))} \leq r\, M(\omega), \quad i = 1, 2, \tag{2.66}$$

for all $2 \leq r < \infty$. If there is another function $u \in C([0,T];\mathcal{H}) \cap L^\infty(0,T;\mathcal{V})$ that satisfies (0.1)–(0.4) for fixed $\omega \in \tilde{\Omega}$, then $\partial(v-u)/\partial t \in L^\infty(0,T;\mathcal{V}')$, where $\mathcal{V}'$ is the dual of $\mathcal{V}$, and we can use the same argument as in [4, pp. 50–51] with the help of (2.66) to find $v \equiv u$. Hence the above v is unique, and ζ is also unique. Thus, ζ is determined independently of any choice of the subsequence $\{\zeta_\epsilon\}$. Next we will show that $\zeta \in C([0,T];L^2(G))$. For this, one could follow the argument in [12], but there seems to be a gap in the argument because of some missing details. We proceed differently. It is already known that $\zeta \in L^\infty(0,T;L^\infty(G)) \cap C([0,T];H^{-\beta}(G))$ for any $\beta \in (0,1/2)$. Thus, $\zeta(t)$ is $L^2(G)$-weakly continuous on $[0,T]$; see Theorem 2.1 of Strauss [11]. Let us recall how ζ was constructed. The approximation ζ_ϵ satisfies the energy identity

$$\|\zeta_\epsilon(t)\|^2_{L^2(G)} = \|\zeta_\epsilon(0)\|^2_{L^2(G)} + 2\int_0^t \int_G F_\epsilon \zeta_\epsilon \, dx\, ds, \tag{2.67}$$

where F_ϵ converges to $-(\nabla^\perp \Psi)\cdot\nabla(\nabla \times W)$ strongly in $C([0,T];L^2(G))$ as $\epsilon \to 0$. We pause to emphasize that a formal computation such as multiplying (2.50) by ζ and integrating over G cannot be justified because ζ is only known to be in $L^\infty((0,T)\times G)$.

For each $t \in [0,T]$, $\zeta_\epsilon(t)$ converges to $\zeta(t)$ strongly in $H^{-\beta}(G)$ for any $\beta \in (0,1/2)$ and $\zeta(t) \in L^2(G)$. Hence, $\zeta_\epsilon(t)$ converges to $\zeta(t)$ weakly in $L^2(G)$. By (2.6), we pass $\epsilon \to 0$ to arrive at

$$\|\zeta(t)\|^2_{L^2(G)} \le \|\zeta(0)\|^2_{L^2(G)} - 2\int_0^t \int_G (\nabla^\perp \Psi)\cdot\nabla(\nabla\times W)\zeta\,dx\,ds, \tag{2.68}$$

from which it follows that

$$\overline{\lim_{t\to 0}}\, \|\zeta(t)\|^2_{L^2(G)} \le \|\zeta(0)\|^2_{L^2(G)}. \tag{2.69}$$

Consequently, we find that $\zeta(t)$ is right-continuous at $t = 0$ strongly in $L^2(G)$. Since $\zeta \in L^\infty(0,T;L^\infty(G)) \cap C([0,T];H^{-\beta}(G))$ for any $\beta \in (0,1/2)$ implies that $\zeta(t) \in L^\infty(G)$ and $v(t) \in \mathcal{V}$ for every $t \in [0,T]$, we can take any $0 < t < T$ as the initial time to repeat the above result. By virtue of the uniqueness of solutions, $\zeta(t)$ is right-continuous at each $t \in [0,T)$ strongly in $L^2(G)$. By change of variables $\zeta \mapsto -\zeta$, $W \mapsto -W$, and $t \mapsto T-t$ and applying the above result, we find that $\zeta(t)$ is also left-continuous, and $\zeta \in C([0,T];L^2(G))$. Thus, $v(t) \in C([0,T];\mathcal{V})$. By the same argument as in the proof of measurability of ζ_ϵ, we can easily find that for each $0 < s \le T$, $\zeta(s,\cdot)$ is $L^2(G)$-valued $\mathcal{F}_s$-measurable. Therefore, $v(s,\cdot)$ is $\mathcal{V}$-valued $\mathcal{F}_s$-measurable. Now the proof of Theorem 1.2 is complete.

3. Proof of Theorem 1.3. We proceed under the assumptions made in section 1.2.2. Let us define

$$W_\epsilon(t,x;\omega) = g_\epsilon(x)\,X_t(\omega), \tag{3.1}$$

where each $g_\epsilon \in C^4(\overline{G})$ and, as $\epsilon \to 0$,

$$g_\epsilon \to g \qquad \text{strongly in } H^1(G). \tag{3.2}$$

Let $\tilde\Omega \in \mathcal{F}_0$ with $P(\tilde\Omega) = 1$ such that for each $\omega \in \tilde\Omega$, $X_{(\cdot)}(\omega) \in C([0,T])$.

As in section 2, we choose $u_{0,\epsilon} \in C^2(\overline{G}) \cap \mathcal{V}$ such that

$$\nabla\times u_{0,\epsilon} \to \nabla\times u_0 \quad \text{strongly in } L^2(G). \tag{3.3}$$

According to Lemma 2.1, for each $\omega \in \tilde\Omega$ there is a unique $\xi_\epsilon(\cdot,\cdot;\omega) \in C\big([0,T];C^1(\overline{G})\big)$ that satisfies

(3.4)
$$\xi_\epsilon(t) = \nabla\times u_{0,\epsilon} - \int_0^t (\nabla^\perp\Psi_\epsilon)\cdot\nabla\xi_\epsilon\,ds + \nabla\times W_\epsilon(t) \qquad \text{for all } (t,x) \in [0,T]\times\overline{G}.$$

Here we have set $\xi_\epsilon = \zeta_\epsilon + \nabla\times W_\epsilon$, where ζ_ϵ satisfies (2.1)–(2.3). Next we need the fact that $H^1(G)$ is a Borel subset of $L^2(G)$. This follows easily from the fact that each closed ball of $H^1(G)$ of finite radius is a closed subset of $L^2(G)$. It is already known that $\xi_\epsilon(s,\cdot)$ is $L^2(G)$-valued $\mathcal{F}_s$-measurable for each $s \in [0,T]$. Since $\xi_\epsilon \in C([0,T];C^1(\overline{G}))$, for each $\omega \in \tilde\Omega$, $\xi_\epsilon(t,\cdot)$ is $H^1(G)$-valued continuous and adapted to $\{\mathcal{F}_t\}$. In the same way, $\nabla^\perp\Psi_\epsilon$ is $H^3(G)$-valued continuous and adapted to $\{\mathcal{F}_t\}$. We now apply Ito's formula to find that, for P-almost all ω,

$$\begin{aligned}\|\xi_\epsilon(t)\|^2_{L^2(G)} = {}& \|\nabla\times u_{0,\epsilon}\|^2_{L^2(G)} - 2\int_0^t \langle(\nabla^\perp\Psi_\epsilon\cdot\nabla)\xi_\epsilon,\,\xi_\epsilon\rangle_{L^2(G)}\,ds \\ &+ 2\int_0^t \langle\xi_\epsilon,\,\nabla\times g_\epsilon\rangle_{L^2(G)}\,dX_s + \int_0^t \|\nabla\times g_\epsilon\|^2_{L^2(G)}\,ds\end{aligned} \tag{3.5}$$

for all $0 \le t \le T$, where the second term in the right-hand side vanishes. By the Burkholder–Davis–Gundy inequality, it holds that, for all $t \in [0, T]$,

$$\begin{aligned} E\biggl(\sup_{0\le\eta\le t}\biggl|\int_0^\eta \langle \xi_\epsilon,\, \nabla\times g_\epsilon\rangle_{L^2(G)}\, dX_s\biggr|\biggr) &\\ \le M\, E\biggl(\biggl(\int_0^t |\langle \xi_\epsilon,\, \nabla\times g_\epsilon\rangle_{L^2(G)}|^2\, ds\biggr)^{1/2}\biggr) &\\ \le M\|\nabla\times g_\epsilon\|_{L^2(G)} E\biggl(\biggl(\int_0^t \|\xi_\epsilon\|^2_{L^2(G)}\, ds\biggr)^{1/2}\biggr). & \end{aligned} \tag{3.6}$$

Here and below, $E(\cdot)$ is the expectation with respect to P, and M denotes positive constants independent of ϵ. We can infer from (3.5) and (3.6) that

$$E\bigl(\|\xi_\epsilon\|^2_{C([0,T];L^2(G))}\bigr) \le M \tag{3.7}$$

and

$$E\bigl(\|\nabla^\perp\Psi_\epsilon\|^2_{C([0,T];H^1(G))}\bigr) \le M. \tag{3.8}$$

Consequently,

$$\begin{aligned} E\biggl(\|(\nabla^\perp\Psi_\epsilon)\cdot\nabla\xi_\epsilon\|_{C([0,T];H^{-1-\delta}(G))}\biggr) &\\ = E\biggl(\|\nabla\cdot((\nabla^\perp\Psi_\epsilon)\xi_\epsilon)\|_{C([0,T];H^{-1-\delta}(G))}\biggr) \le M & \end{aligned} \tag{3.9}$$

for any $\delta \in (0, 1/2)$. From now on, we fix a given $\delta \in (0, 1/2)$. In the meantime, we find from the well-known property of the Brownian motion

$$E\bigl(\|W_\epsilon\|^2_{C^\gamma([0,T];H^1(G))}\bigr) \le M \tag{3.10}$$

for some $0 < \gamma < 1/2$. It now follows from (3.4) and the above estimates that

$$E\bigl(\|\xi_\epsilon\|_{C^\gamma([0,T];H^{-1-\delta}(G))}\bigr) \le M. \tag{3.11}$$

Since it holds that, for all $t_1, t_2 \in [0, T]$ and all $\phi \in C([0, T]; L^2(G)) \cap C^\gamma([0, T]; H^{-1-\delta}(G))$,

$$\begin{aligned} \|\phi(t_2) - \phi(t_1)\|_{H^{-\delta}(G)} &\\ \le M\|\phi(t_2) - \phi(t_1)\|^{1/(1+\delta)}_{L^2(G)}\, \|\phi(t_2) - \phi(t_1)\|^{\delta/(1+\delta)}_{H^{-1-\delta}(G)}, & \end{aligned} \tag{3.12}$$

we can apply the Ascoli theorem to conclude that the injection

$$C([0, T]; L^2(G)) \cap C^\gamma([0, T]; H^{-1-\delta}(G)) \;\to\; C([0, T]; H^{-\delta}(G)) \tag{3.13}$$

is compact. Combining all these, we are now ready to extract a suitably convergent subsequence.

Since the injection

(3.14)
$$L^p\biggl(\Omega; \Bigl[C([0, T]; L^2(G)) \cap C^\gamma([0, T]; H^{-1-\delta}(G))\Bigr]\biggr) \;\to\; L^p\biggl(\Omega; \Bigl[C([0, T]; H^{-\delta}(G))\Bigr]\biggr)$$

is not compact for any $1 \leq p \leq \infty$, we will use a measure-theoretic argument. Let us set $\epsilon = 1/k,\ k = 1, 2, \dots$. It follows from (3.7) and (3.11) that

$$P\biggl(\|\xi_{1/k}\|_{C([0,T];L^2(G))} + \|\xi_{1/k}\|_{C^\gamma([0,T];H^{-1-\delta}(G))} \geq L\biggr) \leq M/L \quad \text{for all } k, \tag{3.15}$$

which yields

$$P\left(\bigcup_{L=1}^{\infty}\bigcap_{m=1}^{\infty}\bigcup_{k=m}^{\infty}\{\|\xi_{1/k}\|_{C([0,T];L^2(G))} + \|\xi_{1/k}\|_{C^\gamma([0,T];H^{-1-\delta}(G))} \leq L\}\right) = 1. \tag{3.16}$$

Hence there is a subset $\Omega^\dagger \subset \tilde{\Omega}$ such that $P(\Omega^\dagger) = 1$ and, for each $\omega \in \Omega^\dagger$, there is a positive integer L and a subsequence $\{\xi_{1/k_j}\}_{j=1}^{\infty}$ satisfying

$$\|\xi_{1/k_j}\|_{C([0,T];L^2(G))} + \|\xi_{1/k_j}\|_{C^\gamma([0,T];H^{-1-\delta}(G))} \leq L \quad \text{for all } j. \tag{3.17}$$

We now fix $\omega \in \Omega^\dagger$, and make a choice of L and a subsequence satisfying (3.17). By virtue of (3.13) and (3.17), we can further extract a subsequence $\{\xi_\nu\}$ such that, as $\nu \to 0$,

$$\xi_\nu \to \xi \quad \text{weak-star in } L^\infty(0,T;L^2(G)), \tag{3.18}$$

$$\xi_\nu \to \xi \quad \text{strongly in } C([0,T];H^{-\delta}(G)) \tag{3.19}$$

for some function ξ. It follows from (2.3) and (3.19) that, as $\nu \to 0$,

$$\Psi_\nu \to \Psi \quad \text{strongly in } C([0,T];H^{2-\delta}(G)), \tag{3.20}$$

where Ψ satisfies

$$\begin{cases} -\Delta\Psi = \xi & \text{in } G, \\ \Psi = 0 & \text{on } \partial G. \end{cases} \tag{3.21}$$

Hence, ξ satisfies, for each $t \in [0,T]$,

$$\xi(t) = \nabla \times u_0 - \int_0^t (\nabla^\perp\Psi)\cdot\nabla\xi\, ds + \nabla \times W \tag{3.22}$$

in the sense of distribution over G. Hence, for each $\omega \in \Omega^\dagger$, we have constructed a function $\xi \in L^\infty(0,T;L^2(G)) \cap C([0,T];H^{-\delta}(G))$ that satisfies (3.22). We will use the following result of Bensoussan and Temam [2, pp. 220–221].

THEOREM. *Let X be a Polish space and Y be a separable Banach space. Suppose that Λ is a multivalued mapping from X to the set of nonempty closed subsets of Y such that its graph is closed. Then, Λ admits a universally Radon measurable selection; i.e., there is a mapping L from X to Y such that $L(x) \in \Lambda(x)$ for all $x \in X$, and L is measurable for any Radon measure defined on the Borel sets of X.*

Since Ω is a Polish space and $\mathcal{F}$ is the completion of $\mathcal{B}(\Omega)$ under P, P is inner regular. Thus, there is a compact subset $K_1 \subset \Omega^\dagger$ such that $P(\Omega^\dagger \setminus K_1) < 1/2$. Let us define a separable Banach space $\mathcal{S} = L^2(0,T;L^2(G)) \cap C\bigl([0,T];H^{-\delta}(G)\bigr)$ and a multivalued mapping

$$\Lambda:\ K_1 \to \mathcal{S}, \tag{3.23}$$

where

$$\Lambda(\omega) = \big\{ \xi \in \mathcal{S} \,\big|\; \xi \text{ satisfies (3.22) with } W = g(x)\,\omega(t) \text{ for all } t \in [0,T] \big\}. \tag{3.24}$$

We will show that the graph of Λ is closed in $K_1 \times \mathcal{S}$. Let $\omega_n \in K_1$ and $\xi_n \in \Lambda(\omega_n)$ for $n = 1, 2, \dots$ such that, as $n \to \infty$,

$$(\omega_n, \xi_n) \;\to\; (\omega^\star, \xi^\star) \quad \text{in } K_1 \times \mathcal{S}. \tag{3.25}$$

For each $n,$ it holds that

$$\xi_n(t) = \nabla \times u_0 - \int_0^t (\nabla^\perp \Psi_n) \cdot \nabla \xi_n \, ds + \nabla \times g(x)\omega_n(t) \tag{3.26}$$

in the sense of distribution over G for all $t \in [0,T]$. Here, Ψ_n is determined from

$$\begin{cases} -\Delta \Psi_n = \xi_n & \text{in } G, \\ \Psi_n = 0 & \text{on } \partial G. \end{cases} \tag{3.27}$$

Since ξ_n converges to $\xi^\star$ in $\mathcal{S},$ and ω_n converges to $\omega^\star$ in $C([0,T]),$ it is apparent that $\xi^\star$ satisfies (3.22) with $\omega^\star$. Hence, $\xi^\star \in \Lambda(\omega^\star)$. By the same argument, the set $\Lambda(\omega)$ is a nonempty closed subset of $\mathcal{S}$ for each $\omega \in K_1$.

Since K_1 is also a Polish space, we apply the above theorem to find a function

$$\Xi_1 \, : K_1 \;\to\; \mathcal{S} \tag{3.28}$$

such that

$$\Xi_1(\omega) \in \Lambda(\omega) \qquad \text{for each } \omega \in K_1 \tag{3.29}$$

and Ξ_1 is measurable for every radon measure defined on the Borel subsets of K_1. Since K_1 is a compact subset of $\Omega,$ it is apparent that

$$\Xi_1^{-1}(\mathcal{O}) \in \mathcal{F} \qquad \text{for every open subset } \mathcal{O} \text{ of } \mathcal{S}. \tag{3.30}$$

Let us proceed by induction. After $K_1, \dots, K_m$ and $\Xi_1, \dots, \Xi_m$ have been chosen, we choose a compact subset $K_{m+1} \subset \Omega^\dagger \backslash \bigcup_{j=1}^m K_j$ such that $P(\Omega^\dagger \backslash \bigcup_{j=1}^{m+1} K_j) < 1/2^{m+1}$ and the corresponding Ξ_{m+1} as above. By piecing up all Ξ_j's, we can construct a $\mathcal{S}$-valued $\mathcal{F}$-measurable function Ξ such that for P-almost all $\omega,$ $\Xi(\omega)$ satisfies, for all $t \in [0,T],$

$$\Xi(t, \cdot\,; \omega) = \nabla \times u_0 - \int_0^t (\nabla^\perp \Psi) \cdot \nabla \Xi \; ds + \nabla \times W \tag{3.31}$$

in the sense of distribution over G, where Ψ is obtained from (3.21) with ξ replaced by Ξ. By setting $v = \nabla^\perp \Psi,$ we repeat the same argument as before to arrive at

$$\frac{\partial v}{\partial t} + (v \cdot \nabla) v + \nabla p = \frac{\partial W}{\partial t} \tag{3.32}$$

in the sense of distribution over $(0,T) \times G$ with some scalar distribution $p,$ for P-almost all $\omega,$ and

$$v(0, \cdot\,; \omega) = u_0 \quad \text{for } P\text{-almost all } \omega. \tag{3.33}$$

From the regularity of Ξ, we have

$$v \in L^2(0,T;\mathcal{V}) \cap C\big([0,T]; H^{1-\delta}(G) \cap \mathcal{H}\big), \quad \text{for } P\text{-almost all } \omega, \tag{3.34}$$

and the measurability of v also follows from that of Ξ. This completes the proof of Theorem 1.3 under the assumptions in section 1.2.2.

It remains to link this special case to a general case. Let $B_t(\cdot)$ be a given standard Brownian motion over $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$. Let $\tilde{\Omega}$ be a subset of Ω with $P(\tilde{\Omega}) = 1$ such that for each $\omega \in \tilde{\Omega}$, $B_{(\cdot)}(\omega) \in C([0,\infty))$. We consider a mapping Θ from $\tilde{\Omega}$ into $C([0,\infty))$ defined by

$$\Theta : \omega \mapsto B_{(\cdot)}(\omega). \tag{3.35}$$

Then Θ is $C([0,\infty))$-valued $\mathcal{F}$-measurable, and Θ induces a probability measure P^* on $\big(C([0,\infty)), \mathcal{B}\big(C([0,\infty))\big)\big)$ such that

$$P^*(\mathcal{G}) = P(\Theta^{-1}(\mathcal{G})) \quad \text{for each } \mathcal{G} \in \mathcal{B}\big(C([0,\infty))\big). \tag{3.36}$$

Then the coordinate mapping process X_t becomes a standard Brownian motion over $\big(C([0,\infty)), \mathcal{B}\big(C([0,\infty))\big), P^*\big)$, and

$$X_t(\Theta(\omega)) = B_t(\omega) \quad \text{for all } \omega \in \tilde{\Omega}. \tag{3.37}$$

We now set $\Omega^* = C([0,\infty))$ and define $\mathcal{F}_t^*$, $\mathcal{F}^*$ by (1.6) and (1.7) with P^*. We note that if $\mathcal{G}$ is a P^*-negligible set, then $\Theta^{-1}(\mathcal{G})$ is P-negligible set and belongs to $\mathcal{F}$ since $\mathcal{F}$ is complete under P. Meanwhile, W is given by

$$W(t,x;\omega) = g(x)B_t(\omega). \tag{3.38}$$

We define

$$W^*(t,x;\omega^*) = g(x)X_t(\omega^*) \tag{3.39}$$

so that, for all $(t,x) \in [0,T] \times G$ and all $\omega \in \tilde{\Omega}$,

$$W^*(t,x;\Theta(\omega)) = W(t,x;\omega). \tag{3.40}$$

Then, we have a random function u^* that is $L^2(0,T;\mathcal{V}) \cap C\big([0,T]; \mathcal{H} \cap H^{1-\delta}(G)\big)$-valued $\mathcal{F}^*$-measurable and satisfies (0.1)–(0.4) with W replaced by W^* for P^*-almost all ω^*. For the solution defined over Ω, we simply set

$$u(t,x;\omega) = u^*(t,x;\Theta(\omega)). \tag{3.41}$$

This completes the proof of Theorem 1.3.

Note added in proof. One can easily show that the last term of (3.6) is finite by using a stopping time in (3.5).

REFERENCES

[1] C. Bardos, *Existence et unicité de la solution de l'équation d'Euler en dimension deux*, J. Math. Anal. Appl., 40 (1972), pp. 769–790.

[2] A. Bensoussan and R. Temam, *Equations stochastiques du type Navier-Stokes*, J. Funct. Anal., 13 (1973), pp. 195–222.

[3] H. Bessaih, *Martingale solutions for stochastic Euler equations*, Stochastic Anal. Appl., 17 (1999), pp. 713–725.

[4] H. Bessaih and F. Flandoli, *2-D Euler equation perturbed by noise*, NoDEA Nonlinear Differential Equations Appl., 6 (1998), pp. 35–54.

[5] Z. Brzezniak and S. Peszat, *Stochastic two dimensional Euler equations*, Ann. Probab., to appear.

[6] M. Capinski and N. Cutland, *Stochastic Euler equations on the torus*, Ann. Appl. Prob., 9 (1998), pp. 688–705.

[7] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed., Springer-Verlag, New York, Berlin, Heidelberg, 1997.

[8] P.L. Lions, *Mathematical Topics in Fluid Mechanics*, Vol. 1, Clarendon Press, Oxford, UK, 1996.

[9] R. Mikulevicius and G. Valiukevicius, *On stochastic Euler equation in R^d*, Electron. J. Probab., 5 (2000), pp. pp. 1–20.

[10] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl., 146 (1987), pp. 65–96.

[11] W.A. Strauss, *On continuity of functions with values in various Banach spaces*, Pacific J. Math., 19 (1966), pp. 543–551.

[12] V.I. Yudovich, *Non-stationary flow of an ideal incompressible liquid*, Zh. Vychisl. Mat. Mat. Fiz., 3 (1963), pp. 1032–1066.

SIAM J. MATH. ANAL.
Vol. 33, No. 5, pp. 1228–1245

EVOLUTION OF ELASTIC CURVES IN $\mathbb{R}^n$: EXISTENCE AND COMPUTATION*

GERHARD DZIUK†, ERNST KUWERT‡, AND REINER SCHÄTZLE§

Abstract. We consider curves in $\mathbb{R}^n$ moving by the gradient flow for elastic energy, i.e., the L^2 integral of curvature. Long-time existence is proved in the two cases when a multiple of length is added to the energy or the length is fixed as a constraint. Along these lines, a lower bound for the lifespan of solutions to the curve diffusion flow is observed. We derive algorithms for both the elastic flows and the curve diffusion equation. After a numerical test we compute several examples, including cases of curve diffusion in which a singularity develops.

Key words. geometric evolution equations, fourth order, elastic curves, algorithms, computations

AMS subject classifications. 53C44, 35K55, 65M60

PII. S0036141001383709

1. Introduction.

In the Bernoulli model of an elastic rod [13] described by a closed curve $f : I = \mathbb{R}/\mathbb{Z} \to \mathbb{R}^n$, the elastic energy is the curvature integral

$$E(f) = \frac{1}{2}\int_I |\varkappa|^2 \, ds. \tag{1.1}$$

Here, putting $\gamma = |\partial_x f|$, we denote by $ds = \gamma\, dx$ the arclength element and by $\partial_s = \gamma^{-1}\partial_x$ the arclength derivative; thus $\tau = \partial_s f$ is the unit tangent and $\varkappa = \partial_s^2 f$ is the curvature vector. The Euler–Lagrange operator for E is given by

$$\operatorname{grad}_{L^2} E(f) = \nabla_s^2 \varkappa + \frac{1}{2}|\varkappa|^2 \varkappa, \tag{1.2}$$

where $\nabla_s \phi$ denotes the normal component of $\partial_s \phi$, i.e., $\nabla_s \phi = \partial_s \phi - \langle \partial_s \phi, \tau \rangle \tau$. Critical points of E subject to fixed length are called *elasticae*. In $\mathbb{R}^3$ the only stable elastica is the simply covered round circle, while in $\mathbb{R}^2$ multiple coverings of the circle and a figure eight solution are also stable [9]. However, there are unstable critical points in $\mathbb{R}^3$ [8].

For any $\lambda > 0$ the elasticae are (up to homothety) exactly the critical points of the energy

$$E_\lambda(f) = E(f) + \lambda\, L(f). \tag{1.3}$$

Thus one may expect that the gradient flow for E_λ, defined by the equation

$$\partial_t f = -\nabla_s^2 \varkappa - \frac{1}{2}|\varkappa|^2 \varkappa + \lambda\, \varkappa, \tag{1.4}$$

*Received by the editors January 15, 2001; accepted for publication (in revised form) August 6, 2001; published electronically April 8, 2002.
http://www.siam.org/journals/sima/33-5/38370.html

†Institut für Angewandte Mathematik, Hermann-Herder-Straße 10, D-79104 Freiburg, Germany (gerd@mathematik.uni-freiburg.de).

‡Mathematisches Institut, Albert-Ludwigs-Universität Freiburg, Eckerstraße 1, D-79104 Freiburg, Germany (kuwert@mathematik.uni-freiburg.de).

§Department Mathematik, ETH Zürich, Rämistraße 101, CH–Zürich, Switzerland (schaetz@math.ethz.ch).

deforms a given curve into an elastica. Alternatively, one may consider the gradient flow for E subject to fixed length. This again leads to (1.4), but now the parameter λ is not constant and depends on the map via

$$\lambda = \frac{\int_I \langle \nabla_s^2 \varkappa + \frac{1}{2}|\varkappa|^2 \varkappa, \varkappa \rangle \, ds}{\int_I |\varkappa|^2 \, ds}. \tag{1.5}$$

For $n = 2$, long-time existence and subconvergence up to translations are due to Polden [10, 11] for (1.4) with $\lambda > 0$ fixed and due to Wen [14] when coupled with (1.5); the result in [14] is obtained only for curves with a nonzero rotation index. Flows approaching elasticae but geometrically different from the above have been investigated by Langer and Singer [9], Koiso [7], and other authors. Here we also briefly consider the equation

$$\partial_t f = -\nabla_s^2 \varkappa; \tag{1.6}$$

for $n = 2$, this is the one-dimensional analogue of surface diffusion [3]. Along the evolution (1.6) the length decreases and, for $n = 2$, the enclosed area is preserved. However, the L^2-norm of $\varkappa$ need not remain bounded, and the solutions may develop singularities [11].

Our analytic results for these evolution problems are presented in section 3. For the curve diffusion flow (1.6), we prove a lower bound on the existence time in terms of the L^2-norm of $\varkappa$, which may be useful when rescaling the flow at a singularity. In case of the elastic flows, where the energy $E(f) = \frac{1}{2}||\varkappa||_{L^2}^2$ is bounded a priori, we show long-time existence and subconvergence up to translations. This generalizes the results of Polden and Wen to an arbitrary dimension and covers the open case of zero rotation number for the length-constrained flow in $\mathbb{R}^2$. Our method, which essentially follows [10, 11, 14], is based on L^2 curvature estimates combined with Gagliardo–Nirenberg-type inequalities. This technique has also been employed by other authors; see for example [1]. For the length-constrained flow, a little extra argument is needed since one of the terms is critical for the interpolation.

The second aspect we address is the problem of numerically computing solutions, guided by ideas which have been developed for the isotropic and anisotropic curve shortening flow in [4, 5]. Based on the variational form of the problems, we shall derive numerical algorithms which will lead to suitable difference schemes with respect to space and time. The schemes are based on a discretization with piecewise linear finite elements for a mixed method. The fourth order problem will be written as a system of second order problems, and so we can avoid the discretization with C^1-elements. The time discretization will be done in a semi-implicit way. The scheme then requires the solution of a system of nearly tridiagonal linear equations in each time step. We provide test computations for the elastic flow in two and three dimensions with and without fixed length of the curve. The curve diffusion case is numerically quite close to the elastic flow problem and so is technically solved in a similar way. We show results for examples with singularities. We do not prove the convergence of the algorithms. This will be a topic of future research.

The graphics were done with the use of the visualization package GRAPE [12].

2. Equations of evolution and inequalities.

We start by considering a time-dependent curve $f : [0, T) \times I \to \mathbb{R}^n$ and put

$$\partial_t f = V + \varphi \, \tau, \tag{2.1}$$

where V is the normal velocity and $\varphi = \langle \partial_t f, \tau \rangle$. As $\langle \partial_s \phi, \tau \rangle = -\langle \phi, \partial_s \tau \rangle = -\langle \phi, \varkappa \rangle$ for any normal field ϕ along f, we have

$$\nabla_s \phi = \partial_s \phi + \langle \phi, \varkappa \rangle \tau \quad (\phi \text{ normal}). \tag{2.2}$$

The first lemma expresses the time derivative of various geometric quantities associated to f in terms of V and φ.

LEMMA 2.1. *The following formulae follow from* (2.1)*:*

$$\partial_t(ds) = (\partial_s \varphi - \langle \varkappa, V \rangle) ds, \tag{2.3}$$

$$\partial_t \partial_s - \partial_s \partial_t = (\langle \varkappa, V \rangle - \partial_s \varphi) \partial_s, \tag{2.4}$$

$$\partial_t \tau = \nabla_s V + \varphi \, \varkappa, \tag{2.5}$$

$$\partial_t \phi = \nabla_t \phi - \langle \nabla_s V + \varphi \varkappa, \phi \rangle \, \tau, \tag{2.6}$$

$$\nabla_t \varkappa = \nabla_s^2 V + \langle \varkappa \, , V \, \rangle \varkappa + \varphi \nabla_s \varkappa, \tag{2.7}$$

$$(\nabla_t \nabla_s - \nabla_s \nabla_t) \phi = (\langle \varkappa, V \rangle - \partial_s \varphi) \nabla_s \phi + \langle \varkappa, \phi \rangle \nabla_s V - \langle \nabla_s V, \phi \rangle \varkappa. \tag{2.8}$$

Here in (2.6) and (2.8) the vectorfield ϕ is assumed to be normal, i.e., $\langle \phi, \tau \rangle \equiv 0$.

Proof. Equations (2.3) and (2.4) follow from

$$\frac{1}{\gamma} \partial_t \gamma = \frac{1}{\gamma} \langle \tau, \, \partial_t \, \partial_x f \rangle = \langle \tau, \, \partial_s (V + \varphi \tau) \rangle = \partial_s \varphi - \langle \varkappa, \, V \rangle.$$

Writing $\tau = \partial_s f$, (2.5) is obtained from (2.4) and (2.2), and (2.6) follows because $\langle \partial_t \phi, \tau \rangle = -\langle \phi, \partial_t \tau \rangle$. The equation for $\varkappa$ combines (2.4) and (2.5), substituting $\varkappa = \partial_s \tau$. Finally, we have, for any normal field ψ using successively (2.2), (2.4), (2.5), and (2.6),

$$\begin{aligned}
\langle \nabla_t \nabla_s \phi, \, \psi \rangle &= \langle \partial_t (\partial_s \phi + \langle \phi, \, \varkappa \rangle \tau), \, \psi \rangle \\
&= \langle \partial_s \, \partial_t \phi, \, \psi \rangle + (\langle \varkappa, \, V \rangle - \partial_s \varphi) \langle \nabla_s \phi, \psi \rangle + \langle \phi, \, \varkappa \rangle \langle \nabla_s V + \varphi \varkappa, \psi \rangle \\
&= \langle \partial_s (\nabla_t \phi - \langle \nabla_s V + \varphi \varkappa, \phi \rangle \tau), \, \psi \rangle \\
&\quad + (\langle \varkappa, \, V \rangle - \partial_s \varphi) \langle \nabla_s \phi, \, \psi \rangle + \langle \phi, \, \varkappa \rangle \langle \nabla_s V + \varphi \varkappa, \psi \rangle \\
&= \langle \nabla_s \nabla_t \phi, \, \psi \rangle + (\langle \varkappa, \, V \rangle - \partial_s \varphi) \langle \nabla_s \phi, \, \psi \rangle \\
&\quad + \langle \varkappa \, , \phi \rangle \langle \nabla_s V, \psi \rangle - \langle \nabla_s V, \phi \rangle \langle \varkappa, \, \psi \rangle,
\end{aligned}$$

which proves formula (2.8). □

In the following, we often use integration by parts for ∇_s, which is possible since

$$\partial_s \langle \phi, \psi \rangle = \langle \nabla_s \phi, \psi \rangle + \langle \phi, \nabla_s \psi \rangle \quad (\phi, \psi \text{ normal}). \tag{2.9}$$

For any (not necessarily normal) variation $f_\varepsilon(x) = f(x) + \varepsilon \phi(x)$, one has, by passing to components as in (2.1) and using (2.3), (2.7), and (2.9),

$$\frac{d}{d\varepsilon} L(f_\varepsilon)|_{\varepsilon=0} = -\int_I \langle \varkappa, \phi \rangle ds, \quad \frac{d}{d\varepsilon} E(f_\varepsilon)|_{\varepsilon=0} = \int_I \left\langle \nabla_s^2 \varkappa + \frac{1}{2} |\varkappa|^2 \varkappa, \, \phi \right\rangle ds.$$

This justifies the formulae (1.2) for the L^2 gradient of E and (1.5) for the Lagrange multiplier in the length-constrained flow. For the curve diffusion (1.6), we observe

$$\frac{d}{dt} L(f) = \int_I \langle \varkappa, \nabla_s^2 \varkappa \rangle ds = -\int_I |\nabla_s \varkappa|^2 ds \leq 0. \tag{2.10}$$

Furthermore, in the $n = 2$ case, the enclosed area will be preserved since, if ν denotes the exterior unit normal,

$$\int_I \langle \partial_t f, \nu\rangle ds = -\int_I \langle \nabla_s^2 \varkappa, \nu\rangle ds = -\int_I \partial_s \langle \nabla_s \varkappa, \nu\rangle ds = 0.$$

Lemma 2.2. *Suppose $f : [0, T) \times I \to \mathbb{R}^n$ moves in a normal direction with velocity $\partial_t f = V$, ϕ is a normal vector field along f, and $\nabla_t \phi + \nabla_s^4 \phi = Y$. Then*

$$\frac{d}{dt}\frac{1}{2}\int_I |\phi|^2 ds + \int_I |\nabla_s^2 \phi|^2 ds = \int_I \langle Y, \phi\rangle ds - \frac{1}{2}\int_I |\phi|^2 \langle \varkappa, V\rangle ds. \tag{2.11}$$

Furthermore $\psi = \nabla_s \phi$ satisfies the equation

$$\nabla_t \psi + \nabla_s^4 \psi = \nabla_s Y + \langle \phi, \varkappa\rangle \nabla_s V - \langle \phi, \nabla_s V\rangle \varkappa + \langle \varkappa, V\rangle \psi. \tag{2.12}$$

Proof. Equation (2.11) follows from the definitions, using (2.3) and (2.9); and (2.12) is a consequence of (2.8). □

Before the next lemma we need to explain some notation. For normal vector fields $\phi_1, \ldots, \phi_k$ along f we denote by $\phi_1 * \cdots * \phi_k$ a term of the type

$$\phi_1 * \ldots * \phi_k = \begin{cases} \langle \phi_{i_1}, \phi_{i_2}\rangle \cdots \langle \phi_{i_{k-1}}, \phi_{i_k}\rangle & \text{for } k \text{ even,} \\ \langle \phi_{i_1}, \phi_{i_2}\rangle \cdots \langle \phi_{i_{k-2}}, \phi_{i_{k-1}}\rangle \phi_{i_k} & \text{for } k \text{ odd,} \end{cases}$$

where $i_1, \ldots, i_k$ is any permutation of $1, \ldots, k$. Slightly more generally, we also allow that some of the ϕ_i are functions, in which case the $*$-product reduces to multiplication. For a normal vectorfield ϕ along f, we denote by $P_\nu^\mu(\phi)$ any linear combination of terms of the type $\nabla_s^{i_1}\phi * \cdots * \nabla_s^{i_\nu}\phi$ with universal, constant coefficients, where $\mu = i_1 + \cdots + i_\nu$ is the total number of derivatives. We observe the two properties:

$$P_\nu^\mu(\phi) * P_\beta^\alpha(\phi) = P_{\nu+\beta}^{\mu+\alpha}(\phi), \quad \nabla_s P_\nu^\mu(\phi) = P_\nu^{\mu+1}(\phi).$$

Lemma 2.3. *Suppose $\partial_t f = -\nabla_s^2 \varkappa + \lambda \varkappa + \sigma |\varkappa|^2 \varkappa$, where $\lambda, \sigma \in \mathbb{R}$. Then for $m \geq 0$ the derivatives of the curvature $\phi_m = \nabla_s^m \varkappa$ satisfy*

$$\nabla_t \phi_m + \nabla_s^4 \phi_m = P_3^{m+2}(\varkappa) + \lambda(\nabla_s^{m+2}\varkappa + P_3^m(\varkappa)) + \sigma(P_3^{m+2}(\varkappa) + P_5^m(\varkappa)).$$

The statement is also true when $\lambda = \lambda(t)$ depends on time.

Proof. For $m = 0$ this follows from (2.7). For $m \geq 1$ we inductively obtain using (2.12)

$$\begin{aligned} \nabla_t \phi_m + \nabla_s^4 \phi_m = \nabla_s &\left[P_3^{m+1}(\varkappa) + \lambda(\nabla_s^{m+1}\varkappa + P_3^{m-1}(\varkappa)) + \sigma(P_3^{m+1}(\varkappa) + P_5^{m-1}(\varkappa))\right] \\ &+ \nabla_s^{m-1}\varkappa * \varkappa * \nabla_s(-\nabla_s^2 \varkappa + \lambda \varkappa + \sigma |\varkappa|^2 \varkappa) \\ &+ \varkappa * (-\nabla_s^2 \varkappa + \lambda \varkappa + \sigma |\varkappa|^2 \varkappa) * \nabla_s^m \varkappa, \end{aligned}$$

and the claim of the lemma follows. □

Next we derive some estimates for curvature integrals, employing as in [10, 11, 14] a variant of the Gagliardo–Nirenberg interpolation inequalities. For this we introduce the scale invariant norms $\|\varkappa\|_{k,p} = \sum_{i=0}^k \|\nabla_s^i \varkappa\|_p$, where

$$\|\nabla_s^i \varkappa\|_p = L(f)^{i+1-\frac{1}{p}} \left(\int_I \left|\nabla_s^i \varkappa\right|^p ds\right)^{\frac{1}{p}}. \tag{2.13}$$

LEMMA 2.4. *Let $f : I \to \mathbb{R}^n$ be a smooth closed curve. Then for any $k \in \mathbb{N}$, $p \geq 2$, and $0 \leq i < k$ we have*

$$\|\nabla_s^i \varkappa\|_p \leq c\, \|\varkappa\|_2^{1-\alpha} \|\varkappa\|_{k,2}^{\alpha}, \tag{2.14}$$

where $\alpha = (i + \frac{1}{2} - \frac{1}{p})/k$ and $c = c(n,k,p)$.

Proof. Assuming $L(f) = 1$ and using the inequality $\big|\partial_s|\phi|\big| \leq |\nabla_s \phi|$ for normal vector fields ϕ which follows from (2.9), the standard proof as in [2] applies. □

PROPOSITION 2.5. *Let f be as in the previous lemma. Then for any term $P_\nu^\mu(\varkappa)$ with $\nu \geq 2$ which contains only derivatives of $\varkappa$ of order at most $k-1$, we have*

$$\int_I |P_\nu^\mu(\varkappa)|\, ds \leq c\, L^{1-\mu-\nu} \|\varkappa\|_2^{\nu-\gamma} \|\varkappa\|_{k,2}^{\gamma}, \tag{2.15}$$

where $\gamma = (\mu + \frac{1}{2}\nu - 1)/k$, and $c = c(n,k,\mu,\nu)$. Moreover if $\mu + \frac{1}{2}\nu < 2k+1$, then $\gamma < 2$ and we have for any $\varepsilon > 0$

$$\int_I |P_\nu^\mu(\varkappa)|\, ds \leq \varepsilon \int_I |\nabla_s^k \varkappa|^2 ds + c\, \varepsilon^{-\frac{\gamma}{2-\gamma}} \left(\int_I |\varkappa|^2\, ds\right)^{\frac{\nu-\gamma}{2-\gamma}} + c \left(\int_I |\varkappa|^2\, ds\right)^{\mu+\nu-1}. \tag{2.16}$$

Proof. By Hölder's inequality and Lemma 2.4 with $p = \nu$, we obtain, if $i_1 + \cdots + i_\nu = \mu$ and $L = L(f)$,

$$\int_I |\nabla_s^{i_1} \varkappa * \cdots * \nabla_s^{i_\nu} \varkappa|\, ds \leq L^{1-\mu-\nu} \prod_{j=1}^{\nu} \|\nabla_s^{i_j} \varkappa\|_\nu \leq c\, L^{1-\mu-\nu} \prod_{j=1}^{\nu} \|\varkappa\|_2^{1-\alpha_j}\, \|\varkappa\|_{k,2}^{\alpha_j}.$$

Here $\alpha_j = (i_j + \frac{1}{2} - \frac{1}{\nu})/k$ and thus $\alpha_1 + \cdots + \alpha_\nu = \gamma$, which proves (2.15). Now a standard interpolation inequality (see [2]) yields

$$\|\varkappa\|_{k,2}^2 \leq c(k)\, (\|\nabla_s^k \varkappa\|_2^2 + \|\varkappa\|_2^2). \tag{2.17}$$

Therefore we obtain, assuming $\gamma < 2$,

$$\begin{aligned} L^{1-\mu-\nu} \|\varkappa\|_2^{\nu-\gamma}\, \|\varkappa\|_{k,2}^{\gamma} &\leq c\, L^{1-\mu-\nu} (\|\varkappa\|_2^{\nu-\gamma}\, \|\nabla_s^k \varkappa\|_2^{\gamma} + \|\varkappa\|_2^{\nu}) \\ &\leq c\, \|\varkappa\|_{L^2}^{\nu-\gamma}\, \|\nabla_s^k \varkappa\|_{L^2}^{\gamma} + c\, L^{1-\mu-\nu/2} \|\varkappa\|_{L^2}^{\nu} \\ &\leq \varepsilon\, \|\nabla_s^k \varkappa\|_{L^2}^2 + c\, \varepsilon^{-\frac{\gamma}{2-\gamma}}\, \|\varkappa\|_{L^2}^{2\, \frac{\nu-\gamma}{2-\gamma}} + c\, L^{1-\mu-\nu} \|\varkappa\|_{L^2}^{\nu}. \end{aligned}$$

Finally, the Poincaré inequality for $\partial_s f$ implies

$$L\, \|\varkappa\|_{L^2}^2 \geq 4\pi^2, \tag{2.18}$$

and inserting this into the last term on the right yields inequality (2.16). □

The following lemma compares the $\nabla_s^m \varkappa$ to the full derivatives $\partial_s^m \varkappa$.

LEMMA 2.6. *We have the identities*

$$\nabla_s \varkappa - \partial_s \varkappa = |\varkappa|^2 \tau, \tag{2.19}$$

$$\nabla_s^m \varkappa - \partial_s^m \varkappa = \sum_{i=1}^{\left[\frac{m}{2}\right]} Q_{2i+1}^{m-2i}(\varkappa) + \sum_{i=1}^{\left[\frac{m+1}{2}\right]} Q_{2i}^{m+1-2i}(\varkappa)\, \tau. \tag{2.20}$$

Here $Q^\mu_\nu(\varkappa)$ denotes a linear combination of terms $\partial_s^{i_1}\varkappa * \cdots * \partial_s^{i_\nu}\varkappa$ with $i_1 + \cdots + i_\nu = \mu$.

Proof. Equation (2.19) is (2.2) for $V = \varkappa$. For $m \geq 2$ we inductively compute

$$\begin{aligned}\nabla_s^m \varkappa - \partial_s^m \varkappa &= \partial_s(\nabla_s^{m-1}\varkappa) + \langle \nabla_s^{m-1}\varkappa, \varkappa\rangle \tau - \partial_s^m \varkappa \\ &= \partial_s \left(\sum_{i=1}^{\left[\frac{m-1}{2}\right]} Q^{m-1-2i}_{2i+1}(\varkappa) + \sum_{i=1}^{\left[\frac{m}{2}\right]} Q^{m-2i}_{2i}(\varkappa)\,\tau \right) \\ &\quad + \langle \partial_s^{m-1}\varkappa, \varkappa\rangle\,\tau + \sum_{i=1}^{\left[\frac{m-1}{2}\right]} \langle Q^{m-1-2i}_{2i+1}(\varkappa), \varkappa\rangle\,\tau,\end{aligned}$$

and the claim follows. □

LEMMA 2.7. *Assume the bounds $\|\varkappa\|_{L^2} \leq \Lambda_0$ and $\|\nabla_s^m \varkappa\|_{L^1} \leq \Lambda_m$ for $m \geq 1$. Then for any $m \geq 1$ one has*

$$\|\partial_s^{m-1}\varkappa\|_{L^\infty} + \|\partial_s^m \varkappa\|_{L^1} \leq c_m(\Lambda_0, \ldots, \Lambda_m). \tag{2.21}$$

Proof. Clearly $\|\partial_s^{m-1}\varkappa\|_{L^\infty} \leq c(n)\,\|\partial_s^m \varkappa\|_{L^1}$, and (2.19) implies $\|\partial_s \varkappa\|_{L^1} \leq \|\nabla_s \varkappa\|_{L^1} + \|\varkappa\|^2_{L^2} \leq \Lambda_1 + \Lambda_0^2$. For $m \geq 2$ we obtain from (2.20)

$$\begin{aligned}\|\partial_s^m \varkappa\|_{L^1} &\leq \|\nabla_s^m \varkappa\|_{L^1} + \sum_{\mu=0}^{m-1} \|Q^\mu_{m+1-\mu}(\varkappa)\|_{L^1} \\ &\leq \|\nabla_s^m \varkappa\|_{L^1} + c\,(\|\varkappa\|_{L^\infty}, \ldots, \|\partial_s^{m-2}\varkappa\|_{L^\infty}) \sum_{\mu=0}^{m-1} \|\partial_s^\mu \varkappa\|_{L^1}.\end{aligned}$$

The claim follows by induction on m. □

3. Estimates and long-time existence.

For the flows considered in this paper, short-time existence is a standard matter, and we only briefly sketch the argument for the case of (1.6). It is sufficient to solve the initial value problem up to a tangential term, i.e., to find a solution $\widetilde{f} : I \times [0, \varepsilon) \longrightarrow \mathbb{R}^n$ of

$$\partial_t \widetilde{f} + \nabla_s^2 \widetilde{\varkappa} = \sigma\, \partial_x \widetilde{f}, \quad \widetilde{f}(0, \cdot\,) = f_0,$$

where $\sigma : [0, \varepsilon) \times I \longrightarrow \mathbb{R}^n$ is an arbitrary function. Namely, in solving the ODE initial value problem $\partial_t \varphi = \sigma(t, \varphi)$ with $\varphi(0, \cdot\,) = id$, one easily verifies that $f(t,x) = \widetilde{f}\big(t, \varphi(t,x)\big)$ satisfies $\partial_t f + \nabla_s^2 \varkappa = 0$. Now from $\partial_s = \frac{1}{\gamma}\partial_x$ and $\nabla_s \phi = \partial_s \phi + \langle \varkappa, \phi\rangle \tau$, one infers

$$\nabla_s^2 \varkappa = \gamma^{-4}(\partial_x^4 f - \langle \partial_x^4 f, \tau\rangle \tau) + \text{ lower order terms.}$$

Thus one can apply standard theory to the equation $\partial_t \widetilde{f} = -\nabla_s^2 \widetilde{\varkappa} + \widetilde{\gamma}^{-4}\langle \partial_x^4 \widetilde{f}, \widetilde{\tau}\rangle \widetilde{\tau}$ and proceed as indicated. The argument applies as it stands to (1.4) when λ is constant; if λ is given by (1.5), one observes that the nonlocal term in the linearization is a compact operator between the relevant parabolic Hölder spaces and argues as before.

We now start with an estimate for the curve diffusion flow (1.6). Recall from [10, 6] that there are examples of finite time singularities in this case.

Theorem 3.1. *Let $f : [0,T) \times I \to \mathbb{R}^n$ be a maximal solution of the curve diffusion equation $\partial_t f = -\nabla_s^2 \varkappa$. If $T < \infty$, then*

$$\int_I |\varkappa|^2 \, ds \geq c\,(T-t)^{-1/4}.$$

Proof. Combining Lemma 2.3 and (2.11) yields for $m \geq 0$

$$\frac{d}{dt}\frac{1}{2}\int_I |\nabla_s^m \varkappa|^2 \, ds + \int_I |\nabla_s^{m+2}\varkappa|^2 \, ds = \int_I \langle P_3^{m+2}(\varkappa), \nabla_s^m \varkappa\rangle \, ds.$$

The terms of type $P_3^{m+2}(\varkappa)$ that contain the $(m+2)$nd derivative have the form $\varkappa * \varkappa * \nabla_s^{m+2}\varkappa$. Integrating by parts, we achieve that only derivatives of order $m+1$ or less occur on the right-hand side. Using (2.16) with $k = m+2$, $\mu = 2m+2$, $\nu = 4$, and $\gamma = 2 - \frac{1}{m+2}$ we have

$$\int_I \langle P_3^{m+2}(\varkappa), \nabla_s^m \varkappa\rangle \, ds \leq \varepsilon \int_I |\nabla_s^{m+2}\varkappa|^2 \, ds + c_m(\varepsilon)\left(\int_I |\varkappa|^2 \, ds\right)^{2m+5}, \tag{3.1}$$

which yields, after absorbing for $\varepsilon = \frac{1}{2}$,

$$\frac{d}{dt}\int_I |\nabla_s^m \varkappa|^2 \, ds + \int_I |\nabla_s^{m+2}\varkappa|^2 \, ds \leq c_m \left(\int_I |\varkappa|^2 \, ds\right)^{2m+5}. \tag{3.2}$$

Now we prove by contradiction that $\limsup_{t \nearrow T} \|\varkappa\|_{L^2} = \infty$. Instead, assuming $\|\varkappa\|_{L^2} \leq \Lambda$ for all $t < T$, we have $\|\nabla_s^m \varkappa\|_{L^2}^2(t) \leq \|\nabla_s^m \varkappa\|_{L^2}^2(0) + c_m(\Lambda)T$ by estimate (3.2). Since $\|\nabla_s^m \varkappa\|_{L^1} \leq L(f)^{\frac{1}{2}} \|\nabla_s^m \varkappa\|_{L^2}$ and $L(f) \leq L(f_0)$ by (2.10), Lemma 2.7 implies

$$\|\partial_s^m \varkappa\|_{L^\infty} \leq c_m(\Lambda, f_0, T) \quad \text{for all } m \in \mathbb{N}_0. \tag{3.3}$$

By the differential equation, we further have

$$\|\partial_s^m V\|_{L^\infty} \leq c_m(\Lambda, f_0, T) \quad \text{for all } m \in \mathbb{N}_0, \quad \|f\|_{L^\infty} \leq c\,(\Lambda, f_0, T). \tag{3.4}$$

Now $\gamma = |\partial_x f|$ satisfies $\partial_t \gamma = -\langle \varkappa, V\rangle\,\gamma$ by (2.3), so that by (3.3) and (3.4) with $m = 0$

$$c^{-1} \leq \gamma \leq c \quad \text{with } c = c\,(\Lambda, f_0, T) > 0. \tag{3.5}$$

For any function $h : I \to \mathbb{R}$ we have $\partial_x^m h - \gamma^m \partial_s^m h = P_m(\gamma, \ldots, \partial_x^{m-1}\gamma, h, \ldots, \partial_s^{m-1}h)$, where P_m is a polynomial. Thus inductively assuming $\|\partial_x^j \gamma\|_{L^\infty} \leq c\,(j, \Lambda, f_0, T)$ for $0 \leq j \leq m-1$, we apply this for $h = \langle \varkappa, V\rangle$ and obtain $\|\partial_x^m \langle \varkappa, V\rangle\|_{L^\infty} \leq c_m(\Lambda, f_0, T)$, using (3.3), (3.4), and (3.5). But differentiating the ODE for γ yields $\partial_t(\partial_x^m \gamma) + \langle \varkappa, V\rangle(\partial_x^m \gamma) \leq c_m(\Lambda, f_0, T)$, which in turn implies

$$\|\partial_x^m \gamma\|_{L^\infty} \leq c_m(\Lambda, f_0, T). \tag{3.6}$$

Then by (3.4), from the equations $|\partial_s f| = 1$ and $\partial_s^2 f = \varkappa$ and from the estimates (3.3) and (3.6), we conclude $\|\partial_x^m f\|_{L^\infty} \leq c_m(\Lambda, f_0, T)$. Together with (3.4) and (3.5), this means that f extends smoothly to $[0,T] \times I$, and in fact even beyond T by short-time existence, which contradicts the maximality of T and therefore proves that $\varkappa$ cannot be uniformly bounded in L^2 on $[0,T)$. Now (3.2) for $m=0$ means

$$\frac{d}{dt}\int_I |\varkappa|^2\, ds \le c\left(\int_I |\varkappa|^2\, ds\right)^5,$$

and the theorem follows by integrating on $[t,t_l]$, where $t_l \nearrow T$ with $\|\varkappa\|_{L^2} \to \infty$. □

We next turn our attention to the elastic flows where the L^2 integral of $\varkappa$ is bounded a priori.

THEOREM 3.2. *For any $\lambda \in [0,\infty)$ and smooth initial data f_0, the L^2 gradient flow* (1.4) *for $E_\lambda(f) = \int_I \left(\frac{1}{2}|\varkappa|^2 + \lambda\right) ds$ has a global solution. If $\lambda > 0$, then as $t_i \to \infty$ the curves $f(t_i,\cdot)$ subconverge, when reparametrized by arclength and suitably translated, to an elastica.*

Proof. By (2.11) and Lemma 2.3 we have

$$\tag{3.7} \begin{aligned} &\frac{d}{dt}\frac{1}{2}\int_I |\nabla_s^m \varkappa|^2\, ds + \int_I |\nabla_s^{m+2}\varkappa|^2\, ds + \lambda\int_I |\nabla_s^{m+1}\varkappa|^2\, ds \\ &\quad = \lambda\int_I \langle P_3^m(\varkappa), \nabla_s^m\varkappa\rangle ds + \int_I \langle P_3^{m+2}(\varkappa) + P_5^m(\varkappa), \nabla_s^m\varkappa\rangle ds. \end{aligned}$$

Recalling (3.1) and using (2.16) for $k=m+2$, $\mu = 2m$, and $\nu = 6$, the last integral is estimated under an assumed bound $E(f) \le \Lambda$ by

$$\tag{3.8} \int_I \langle P_3^{m+2}(\varkappa) + P_5^m(\varkappa), \nabla_s^m\varkappa\rangle ds \le \varepsilon\int_I |\nabla_s^{m+2}\varkappa|^2 ds + c_m(\Lambda,\varepsilon).$$

Again by (2.16) but now for $k=m+2$, $\mu=2m$, and $\nu=4$ we infer

$$\int_I \langle P_3^m(\varkappa), \nabla_s^m\varkappa\rangle ds \le \varepsilon\int_I |\nabla_s^{m+2}\varkappa|^2 ds + c_m(\Lambda)\,(\varepsilon^{-\frac{2m+1}{3}} + 1).$$

Multiplying by λ and replacing ε with $\varepsilon/|\lambda|$ yield

$$\tag{3.9} \lambda\int_I \langle P_3^m(\varkappa), \nabla_s^m\varkappa\rangle ds \le \varepsilon\int_I |\nabla_s^{m+2}\varkappa|^2 ds + c_m(\Lambda,\varepsilon)\,(|\lambda|^{\frac{2(m+2)}{3}} + 1).$$

Now if $E_\lambda(f_0) \le \Lambda$, we obtain, by combining (3.7), (3.8), and (3.9),

$$\tag{3.10} \frac{d}{dt}\int_I |\nabla_s^m\varkappa|^2 ds + \int_I |\nabla_s^{m+2}\varkappa|^2 ds \le c_m(\lambda,\Lambda).$$

On the other hand (2.3) yields

$$\frac{d}{dt}L(f) + \int_I |\nabla_s\varkappa|^2 ds + \lambda\int_I |\varkappa|^2 ds = \frac{1}{2}\int_I |\varkappa|^4 ds,$$

and (2.16) with $k=1$, $\mu=0$, and $\nu=4$ implies

$$\tag{3.11} \frac{d}{dt}L(f) + \frac{1}{2}\int_I |\nabla_s\varkappa|^2 ds \le c\,(\Lambda).$$

Together with (3.10) and (3.11) the argument in the proof of Theorem 3.1 yields long-time existence. Now if $\lambda > 0$, then by (2.18) and the energy bound we have a length bound

$$2\pi^2/\Lambda \leq L(f) \leq \Lambda/\lambda. \tag{3.12}$$

Taking $k = m+2$ in (2.17) and inserting into (3.10) therefore imply

$$\frac{d}{dt}\int_I |\nabla_s^m \varkappa|^2 ds + c_0 \int_I |\nabla_s^m \varkappa|^2 ds \leq c_m(\lambda, \Lambda),$$

where $c_0 = c_0(\lambda, \Lambda) > 0$. This yields a bound $\|\nabla_s^m \varkappa\|^2_{L^2}(t) \leq \|\nabla_s^m \varkappa\|^2_{L^2}(0) + c_m(\lambda, \Lambda)$, and by (3.12) and Lemma 2.7 one concludes

$$\|\partial_s^m \varkappa\|_{L^\infty} + \|\nabla_s^m \varkappa\|_{L^\infty} \leq c_m(\lambda, f_0). \tag{3.13}$$

Thus if $\widetilde{f}(t,\cdot)$ is the reparametrization of $f(t,\cdot)$ by arclength, then as $t \to \infty$, subsequences $\widetilde{f}(t_i,\cdot) - p_i$ converge smoothly to a limit curve f_∞ for an appropriate choice of p_i. Lemma 2.3 and the estimate (3.13) imply $\|\nabla_t(\nabla_s^m \varkappa)\|_{L^\infty} \leq c_m(\lambda, f_0)$. From this and (3.12), (3.13) one sees that the function $u(t) = \|V\|^2_{L^2}(t)$ satisfies $|\dot{u}(t)| \leq c(\lambda, f_0)$, where on the other hand $u \in L^1((0,\infty))$ by the energy identity. Therefore $u(t) \to 0$ as $t \to \infty$ which means that f_∞ is an elastica. □

As one readily checks, the interpolation argument breaks down exactly for a quintic term $|\varkappa|^4\varkappa$ on the right-hand side of the evolution equation for f, since then equality holds in the condition $\mu + \frac{1}{2}\nu < 2k+1$ of Proposition 2.5 as $\mu = 2m+2$, $\nu = 6$, and $k = m+2$. Let us finally consider the length-constrained flow.

THEOREM 3.3. *The gradient flow for $E(f) = \frac{1}{2}\int_I |\varkappa|^2 ds$ subject to fixed length $L(f) = L_0$ has a global solution for any smooth initial curve f_0. As $t \to \infty$, the curves subconverge, after reparametrization by arclength and translation, to an elastica.*

Proof. One again has (3.7) where now $\lambda = \lambda(t)$ is given by (1.5). Choosing Λ with $E(f_0) + L(f_0) \leq \Lambda$, we estimate using (2.18) and (2.16) with $k = 1$, $\mu = 0$, and $\nu = 4$

$$|\lambda| \leq c(\Lambda)\,(\|\nabla_s \varkappa\|^2_{L^2} + \|\varkappa\|^4_{L^4}) \leq c(\Lambda)(\|\nabla_s \varkappa\|^2_{L^2} + 1).$$

Using (2.14) with $k = m+2$, $p = 2$, $i = 1$, and $\alpha = \frac{1}{m+2}$, followed by (2.17) and Young's inequality, we obtain

$$|\lambda| \leq c_m(\Lambda)\,(\|\nabla_s^{m+2}\varkappa\|^{\frac{2}{m+2}}_{L^2} + 1).$$

Combination with (3.9) further implies, as $\frac{2}{m+2} \cdot \frac{2(m+2)}{3} = \frac{4}{3} < 2$,

$$\lambda \int_I \langle P_3^m(\varkappa), \nabla_s^m \varkappa\rangle\, ds \leq \varepsilon \int_I |\nabla_s^{m+2}\varkappa|^2 ds + c_m(\Lambda, \varepsilon). \tag{3.14}$$

As to the other term in (3.7) which contains λ, note that (2.14) only gives an inequality $\|\nabla_s^{m+1}\varkappa\|_{L^2} \leq c_m(\Lambda)\|\varkappa\|^\alpha_{m+2,2}$ where $\alpha = \frac{m+1}{m+2}$, which means that the term has critical scaling and the interpolation technique used up to now fails. Instead we employ the scaling properties of E and L: under a dilation $f \to f_\alpha = p + \alpha(f-p) = f + (\alpha - 1)(f-p)$ centered at some point $p \in \mathbb{R}^n$, the energy multiplies by $1/\alpha$ while the length goes like α. Taking the derivative at $\alpha = 1$ and using the definition (1.4) of the gradient flow we see that

$$E(f) - \lambda\, L(f) = -\frac{d}{d\alpha}\left(E(f_\alpha) + \lambda\, L(f_\alpha)\right)|_{\alpha=1} = \int_I \langle \partial_t f, f - p\rangle\, ds.$$

Since $|f - p| \leq L_0$ for appropriate $p = p(t)$, for example taking $p(t) = \int_I f\, ds/L$, this implies $-\lambda \leq L_0^{1/2} \|\partial_t f\|_{L^2}$. For $\lambda^-(t) = -\min\{\lambda(t), 0\}$, one thus gets from the energy identity the estimate

$$\int_0^t \lambda^-(\tau)^2 d\tau \leq c(\Lambda). \tag{3.15}$$

Now the integral in (3.7) is bounded by

$$-\lambda \int_I |\nabla_s^{m+1} \varkappa|^2 ds \leq \varepsilon \int_I |\nabla_s^{m+2} \varkappa|^2 ds + c(\varepsilon)\, \lambda^-(t)^2 \int_I |\nabla_s^m \varkappa|^2 ds. \tag{3.16}$$

From (3.7), (3.8), (3.14), and (3.16) we conclude

$$\frac{d}{dt} \int_I |\nabla_s^m \varkappa|^2\, ds + c_0 \int_I |\nabla_s^m \varkappa|^2\, ds \leq c_m(\Lambda) \left(1 + \lambda^-(t)^2 \int_I |\nabla_s^m \varkappa|^2\, ds\right),$$

where $c_0 = c_0(\Lambda) > 0$ and the Poincaré inequality (2.17) was used. Defining the function $u_m(t) = \exp(c_0 t)\, \|\nabla_s^m \varkappa\|_{L^2}^2$, we have $\dot{u}_m(t) \leq c_m(\Lambda)\, [\, \exp(c_0 t) + \lambda^-(t)^2\, u_m(t)\,]$, and the Gronwall lemma implies

$$u_m(t) \leq e^{a_m(t)} \left(u_m(0) + c_m(\Lambda) \int_0^t e^{c_0 \tau} d\tau\right).$$

Here $a_m(t) = c_m(\Lambda) \int_0^t \lambda^-(\tau)^2\, d\tau \leq c_m(\Lambda)$ by (3.15), and we finally obtain

$$\|\nabla_s^m \varkappa\|_{L^2}^2 \leq c_m(\Lambda) \left(1 + e^{-c_0 t} \|\nabla_s^m \varkappa\|_{L^2}^2(0)\right) \leq c_m(f_0);$$

in particular $|\lambda| \leq c_m(f_0)$. From here the proof proceeds as in Theorem 3.2. □

4. Numerical algorithm.

The numerical algorithm we propose is based on the fact that the evolution equations have a weak formulation. In fact (2.2) implies

$$\nabla_s^2 \varkappa + \frac{1}{2} |\varkappa|^2 \varkappa = \partial_s \left(\partial_s \varkappa + \frac{3}{2} |\varkappa|^2 \tau\right), \tag{4.1}$$

which reveals the divergence form of the main part. In order to avoid C^1-elements for the discretization of (1.4), we rewrite that equation as a system of second order problems for the position vector f and the curvature vector $\varkappa$. It is important to use the mean curvature vector and not the curvature as an additional variable, because we shall work with piecewise affine functions, and the discrete position vector then will be a parametrization of a polygon with piecewise constant and discontinuous normal.

We shall describe the development of the numerical algorithm for the main problem (1.4). It is equivalent to the system

$$\partial_t f + \partial_s^2 \varkappa + \frac{3}{2} \partial_s (|\varkappa|^2 \partial_s f) - \lambda \varkappa = 0, \tag{4.2}$$

$$\varkappa - \partial_s^2 f = 0. \tag{4.3}$$

Because of $\partial_s = \partial_x / |\partial_x f|$, a variational formulation of (4.2), (4.3) is given by

$$\int_I \partial_t f |\partial_x f| \varphi \, dx - \int_I \frac{\partial_x \varkappa}{|\partial_x f|} \partial_x \varphi \, dx - \frac{3}{2} \int_I |\varkappa|^2 \frac{\partial_x f}{|\partial_x f|} \partial_x \varphi \, dx - \lambda \int_I \varkappa \varphi |\partial_x f| \, dx = 0, \tag{4.4}$$

$$\int_I \varkappa |\partial_x f| \psi \, dx + \int_I \frac{\partial_x f}{|\partial_x f|} \partial_x \psi \, dx = 0 \tag{4.5}$$

for all test functions $\varphi, \psi \in H^1(\mathbb{R}/\mathbb{Z}, \mathbb{R}^n)$. We use this weak form of our problem for a finite element discretization in space, which in one space dimension will lead to a suitable difference scheme. Let $I = \bigcup_{j=1}^N I_j$ be a decomposition of $I = \mathbb{R}/\mathbb{Z}$ into intervals given by the nodes x_j, $I_j = [x_{j-1}, x_j]$. Let $h_j = |I_j|$ and $h = \max_{j=1,\dots,N} h_j$ be the maximal diameter of a grid element. For a discretization of (4.4), (4.5) we replace the continuous space $X = H^1(\mathbb{R}/\mathbb{Z}, \mathbb{R}^n)$ by the discrete finite-dimensional space

$$X_h = \left\{ w \in C^0(\mathbb{R}/\mathbb{Z}, \mathbb{R}^n) | \; w|_{I_j} \in \mathbb{P}_1(I_j), j = 1, \dots, N \right\} = Y_h^n$$

of continuous (periodic) piecewise affine functions on the grid. Here Y_h is the space of scalar piecewise affine functions. We use the scalar nodal basis functions $\varphi_j \in Y_h$, $\varphi_j(x_i) = \delta_{ij}$.

$$X_h = Y_h^n = \left(\operatorname{span}\{\varphi_1, \dots, \varphi_N\} \right)^n.$$

We shall use the pointwise interpolation $I_h w$ of a suitable function $w \in C^0(\mathbb{R}/\mathbb{Z}, \mathbb{R}^n)$ which is uniquely defined by $I_h w \in X_h$ and $I_h w(x_j) = w(x_j)$ $(j = 1, \dots, N)$.

A spatially discrete solution of (4.4), (4.5) will be a pair of functions $f_h : [0,T] \to X_h$, $\varkappa_h : [0,T] \to X_h$,

$$f_h(x,t) = \sum_{j=1}^N f_j(t) \varphi_j(x), \quad \varkappa_h(x,t) = \sum_{j=1}^N \varkappa_j(t) \varphi_j(x).$$

Note that each f_j or $\varkappa_j$ is a vector in $\mathbb{R}^n$. It is also worth noticing that $\varkappa_h$ cannot be the second derivative of the piecewise linear position vector f_h.

We look for $f_h(\cdot, t), \varkappa_h(\cdot, t) \in X_h$, $t \in [0,T]$, such that

$$f_h(\cdot, 0) = f_{h0} = I_h f_0$$

and, for all discrete test functions $\varphi_h, \psi_h \in X_h$,

$$\begin{aligned} \int_I I_h \left(\partial_t f_h \varphi_h \right) |\partial_x f_h| \, dx - \int_I \frac{\partial_x \varkappa_h}{|\partial_x f_h|} \partial_x \varphi_h \, dx - \frac{3}{2} \int_I |\varkappa_h|^2 \frac{\partial_x f_h}{|\partial_x f_h|} \partial_x \varphi_h \, dx \\ - \lambda \int_I I_h \left(\varkappa_h \varphi_h \right) |\partial_x f_h| \, dx = 0, \end{aligned} \tag{4.6}$$

$$\int_I I_h \left(\varkappa_h \psi_h \right) |\partial_x f_h| \, dx + \int_I \frac{\partial_x f_h}{|\partial_x f_h|} \partial_x \psi_h \, dx = 0. \tag{4.7}$$

We have used the lumping of masses for practical reasons in both equations. The first equation holds for all $t \in (0,T]$ and the second for all $t \in [0,T]$, and so gives the initial data for $\varkappa_h$. We insert $\varphi_h = \varphi_j$, and $\psi_h = \varphi_j$ $(j = 1, \ldots, N)$ separately for each component into these equations and integrate. The discrete weak equations (4.6), (4.7) are equivalent to the following system of nN ODEs:

$$\begin{aligned} &\frac{1}{2}\left(|f_j - f_{j-1}| + |f_{j+1} - f_j|\right)(\partial_t f_j - \lambda\varkappa_j) - \left(\frac{\varkappa_j - \varkappa_{j-1}}{|f_j - f_{j-1}|} - \frac{\varkappa_{j+1} - \varkappa_j}{|f_{j+1} - f_j|}\right) \\ &\qquad - \frac{1}{2}\left(|\varkappa_{j-1}|^2 + \varkappa_{j-1}\varkappa_j + |\varkappa_j|^2\right)\frac{f_j - f_{j-1}}{|f_j - f_{j-1}|} \\ &\qquad + \frac{1}{2}\left(|\varkappa_j|^2 + \varkappa_j\varkappa_{j+1} + |\varkappa_{j+1}|^2\right)\frac{f_{j+1} - f_j}{|f_{j+1} - f_j|} = 0, \end{aligned} \tag{4.8}$$

$$\frac{1}{2}\left(|f_j - f_{j-1}| + |f_{j+1} - f_j|\right)\varkappa_j + \frac{f_j - f_{j-1}}{|f_j - f_{j-1}|} - \frac{f_{j+1} - f_j}{|f_{j+1} - f_j|} = 0 \tag{4.9}$$

$(j = 1, \ldots, N)$, where $f_0 = f_N, f_{N+1} = f_1, \varkappa_0 = \varkappa_N, \varkappa_{N+1} = \varkappa_1$, and the initial values are given by $f_j(0) = f_0(x_j)$ $(j = 1, \ldots, N)$.

We discretize the scheme (4.8), (4.9) with respect to time in a semi-implicit way, which is similar to the time discretization used for the curve shortening flow in isotropic or anisotropic form in [4, 5]. We use the notation

$$w^m = w(\cdot, m\tau)$$

for the evaluation of a generic function on the mth time level. τ is the chosen time step size with $\tau M = T$. Let us formulate the algorithm for the elastic flow with fixed parameter λ for curves in $\mathbb{R}^n$.

Algorithm 1 (elastic flow with fixed parameter). *For $j = 1, \ldots, N$ let $f_j^0 = f_0(x_j)$, $h_j^0 = |f_j^0 - f_{j-1}^0|$, and*

$$\varkappa_j^0 = \frac{2}{h_{j+1}^0(h_j^0 + h_{j+1}^0)} f_{j+1}^0 - \frac{2}{h_j^0 h_{j+1}^0} f_j^0 + \frac{2}{h_j^0(h_j^0 + h_{j+1})} f_{j-1}^0.$$

For $m = 0, \ldots, M-1$ compute the quantities

$$\begin{aligned} h_j^m &= |f_j^m - f_{j-1}^m|, \\ \beta_j^m &= |\varkappa_{j-1}^m|^2 + \varkappa_{j-1}^m \varkappa_j^m + |\varkappa_j^m|^2, \\ \lambda_j^m &= \lambda, \end{aligned} \tag{4.10}$$

and determine $f_j^{m+1} \in \mathbb{R}^n$ and $\varkappa_j^{m+1} \in \mathbb{R}^n$ $(j = 1, \ldots, N)$ from the linear system

$$\begin{aligned} &\frac{\beta_j^m}{2h_j^m} f_{j-1}^{m+1} + \left(\frac{h_j^m + h_{j+1}^m}{2\tau} - \frac{\beta_j^m}{2h_j^m} - \frac{\beta_{j+1}^m}{2h_{j+1}^m}\right) f_j^{m+1} + \frac{\beta_{j+1}^m}{2h_{j+1}^m} f_{j+1}^{m+1} \\ &\quad + \frac{1}{h_j^m}\varkappa_{j-1}^{m+1} - \left(\frac{1}{h_j^m} + \frac{1}{h_{j+1}^m} + \frac{\lambda_j^m}{2}(h_j^m + h_{j+1}^m)\right)\varkappa_j^{m+1} + \frac{1}{h_{j+1}^m}\varkappa_{j+1}^{m+1} \\ &\qquad = \frac{h_j^m + h_{j+1}^m}{2\tau} f_j^m, \end{aligned} \tag{4.11}$$

$$\frac{1}{2}(h_j^m + h_{j+1}^m)\varkappa_j^{m+1} - \frac{1}{h_j^m} f_{j-1}^{m+1} + \left(\frac{1}{h_j^m} + \frac{1}{h_{j+1}^m}\right) f_j^{m+1} - \frac{1}{h_{j+1}^m} f_{j+1}^{m+1} = 0 \tag{4.12}$$

($j = 1, \dots, N$) with the periodic boundary conditions $f_{i+N} = f_i$, $\varkappa_{i+N} = \varkappa_i$ ($i \in \mathbb{Z}$).

The algorithm for elastic flow with fixed length requires a slight change of the above algorithm. We have to compute the parameter λ in each time step according to a discrete version of (1.5). From the time continuous weak form (4.6) with $\varphi_h = \varkappa_h$ and (4.6) with $\psi_h = f_{ht}$, we can compute the (time-dependent) parameter λ as follows:

$$\begin{aligned} 0 = \frac{d}{dt}\int_I |f_{hx}|\,dx &= \int_I \frac{f_{hx}}{|f_{hx}|} f_{hxt}\,dx = -\int_I I_h(f_{ht}\varkappa_h)|f_{hx}|\,dx \\ &= -\int_I \frac{|\varkappa_{hx}|^2}{|f_{hx}|} - \frac{3}{2}\int_I |\varkappa_h|^2 \frac{f_{hx}}{|f_{hx}|}\varkappa_{hx}\,dx - \lambda\int_I I_h(|\varkappa_h|^2)|f_{hx}|\,dx. \end{aligned}$$

This gives

$$\lambda = -\left(\int_I \frac{|\varkappa_{hx}|^2}{|f_{hx}|}\,dx + \frac{3}{2}\int_I |\varkappa_h|^2\frac{f_{hx}}{|f_{hx}|}\varkappa_{hx}\,dx\right)\bigg/\int_I I_h(|\varkappa_h|^2)|f_{hx}|\,dx.$$

After evaluation of the integrals, we can formulate the algorithm for the length preserving elastic flow. Because of the explicit time discretization of λ, the length of the polygon is conserved only up to an additional error of size τ.

Algorithm 2 (length preserving elastic flow). *Replace* (4.10) *in Algorithm* 1 *with*

$$\lambda_j^m = -3\left(\sum_{i=1}^N \frac{|\varkappa_i^m - \varkappa_{i-1}^m|^2}{h_i^m} + \frac{1}{2}\sum_{i=1}^N (f_i^m - f_{i-1}^m)(\varkappa_i^m - \varkappa_{i-1}^m)\frac{\beta_i^m}{h_i^m}\right)\bigg/\sum_{i=1}^N h_i^m\beta_i^m.$$

The curve diffusion problem (1.6) is equivalent to the system

$$\begin{aligned} \partial_t f + \partial_s^2\varkappa + \frac{3}{2}\partial_s(|\varkappa|^2\partial_s f) - \frac{1}{2}|\varkappa|^2\varkappa = 0, \\ \varkappa - \partial_s^2 f = 0. \end{aligned}$$

The discretization procedure is analogous to the previous ones for the elastic flow with fixed parameter and for elastic flow with fixed length. We just give the algorithm here.

Algorithm 3 (curve diffusion). *Replace* (4.10) *in Algorithm* 1 *with*

$$\lambda_j^m = \frac{1}{2}|\varkappa_j^m|^2.$$

Obviously the discrete system (4.11), (4.12) does not contain the grid parameter h and thus is an intrinsic algorithm for an intrinsic problem. The grid size of the parametrization enters only via the initial condition for f_h. The linear system (4.10), (4.11), (4.12) has the form

$$\left(\frac{1}{\tau}D_m + A_m(\beta_m)\right)f^{m+1} - (A_m + \lambda D_m)\varkappa^{m+1} = \frac{1}{\tau}D_m f^m, \tag{4.13}$$

$$D_m\varkappa^{m+1} + A_m f^{m+1} = 0, \tag{4.14}$$

where f^{m+1} stands for the N-vector which consists of the kth components ($k \in \{1, \dots, n\}$) of f_j^{m+1} ($j = 1, \dots, N$), and similarly for $\varkappa^{m+1}$. This means that the system decouples with respect to the components of the position vector and of the

Table 4.1
Absolute errors and experimental orders of convergence for the test problem with $\tau = 0.5h^2$.

N	h	$\|f - f_h\|_{L^\infty}$	*eoc*	$\|\varkappa - \varkappa_h\|_{L^\infty}$	*eoc*
10	6.2832e-1	1.8943e-2	-	3.6890e-2	-
20	3.1416e-1	3.8717e-3	2.29	8.8264e-3	2.06
40	1.5708e-1	9.2403e-4	2.07	2.1864e-3	2.01
80	7.8540e-2	2.2837e-4	2.02	5.4654e-4	2.00
160	3.9270e-2	5.7121e-5	2.00	1.3665e-4	2.00
320	1.9635e-2	1.4388e-5	1.99	3.4160e-5	2.00
640	9.8175e-3	3.7278e-6	1.95	8.6187e-6	1.99

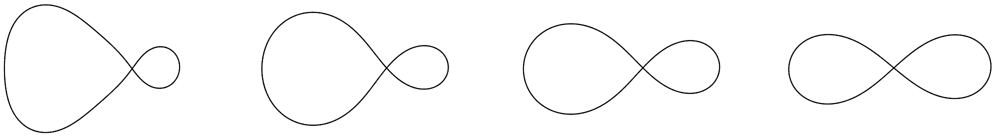


Fig. 4.1. *Length preserving elastic flow for a nonsymmetric lemniscate.*

curvature vector. D_m is a diagonal matrix and $A_m(\beta_m)$ and A_m are tridiagonal matrices depending on the results of the previous time step. Because of the periodicity assumption, the tridiagonality is true only up to two entries in the upper right and lower left corner of the matrix. But this is not a difficulty for the implementation of the linear solver.

We performed test computations for a problem with $\lambda = 0$, for which we know the exact solution of the elastic flow problem. The exact solution is given by

$$f(x,t) = (1+2t)^{\frac{1}{4}}\,(\cos g(x), \sin g(x)), \;\; \varkappa(x,t) = -\,(1+2t)^{-\frac{1}{4}}\,(\cos g(x), \sin g(x)),$$

and we have chosen $g(x) = x + 0.1 \sin x$ in order to create a nonsymmetric distribution of the nodes. We used a uniform grid in the parameter domain I. In Table 4.1 we give the errors in $L^\infty((0,1) \times I)$ between the continuous and the discrete position vector and between the continuous and the discrete curvature vector. We have used $\tau = 0.5h^2$ for the time step size. The experimental order of convergence between two successive grids with grid sizes h_1 and h_2 with respect to the error $E(h)$ is defined by

$$eoc = \log \frac{E(h_1)}{E(h_2)} \Big/ \log \frac{h_1}{h_2}.$$

The results show convergence of second order for the position vector and for the curvature vector in the maximum norm. Let us mention that for the choice $\tau = 0.5h$ one obtains linear convergence.

Figures 4.1 and 4.2 show the evolution of the same initial nonsymmetric lemniscate under length preserving elastic flow and under curve diffusion. Elastic flow converges to a symmetric figure eight curve as a stationary solution while curve diffusion produces a singularity in finite time. Curve diffusion reduces the length of the curve but keeps the enclosed (signed) area constant. The enclosed area in Figure 4.4 is zero, and this leads to a singularity in finite time. The curve “disappears.”

In Figure 4.3 we show results of the computation of the evolution of a plane curve under the length preserving elastic flow (1.4).

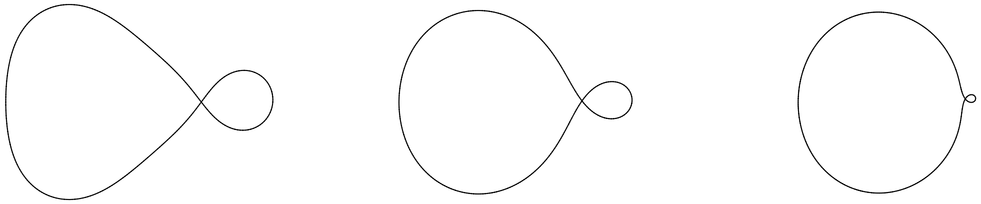


FIG. 4.2. *Curve diffusion for a nonsymmetric lemniscate; development of a singularity.*

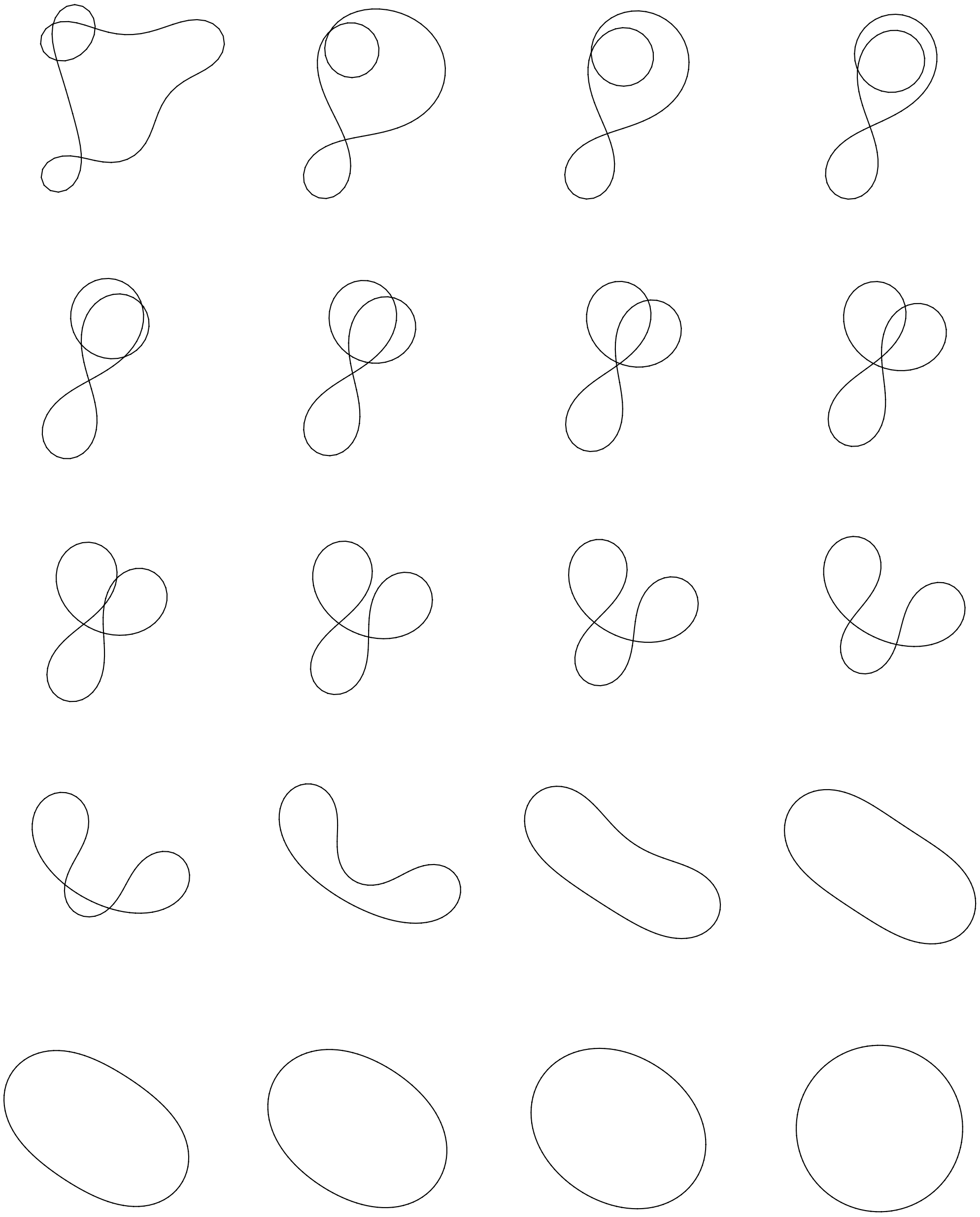


FIG. 4.3. *Time series of the two-dimensional length preserving elastic flow (scaled).*

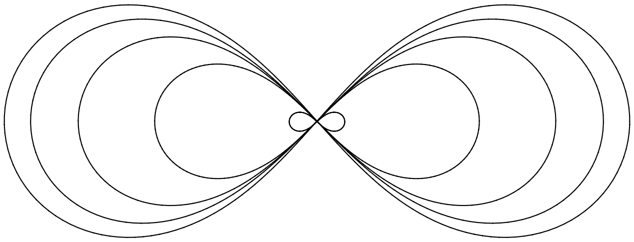


FIG. 4.4. *Curve diffusion for a symmetric lemniscate; development of a singularity.*

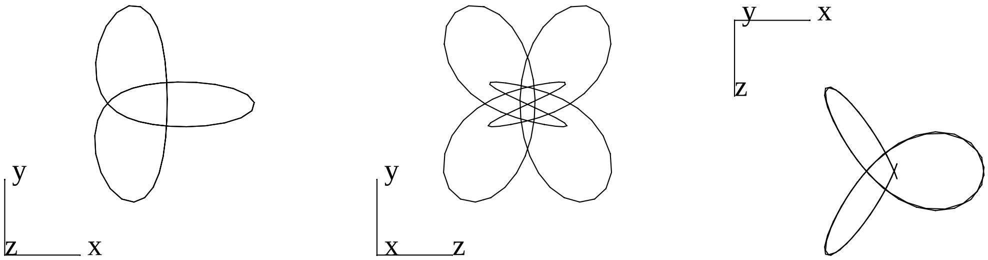


FIG. 4.5. *Three projections of the initial space curve.*

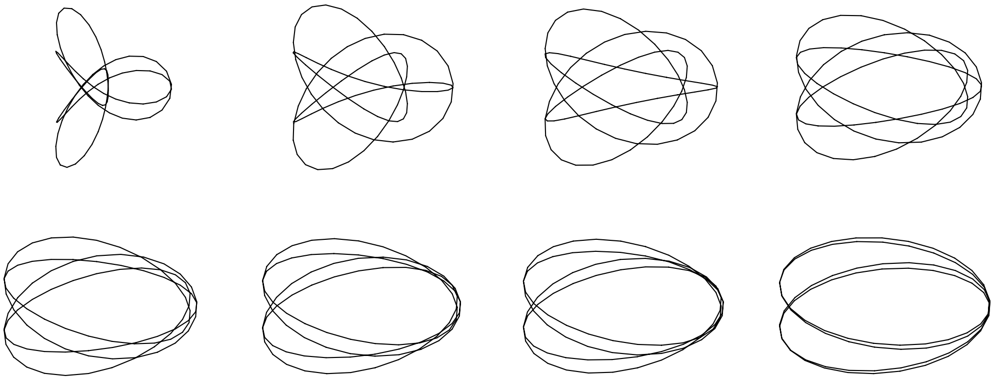


FIG. 4.6. *Elastic flow of the space curve from Figure* 4.5. *From the initial knot to two double circles:* $t = 0.0$, 0.0197, 0.0394, 0.0788, 0.118, 0.158, 0.177, 0.394.

An example for the elastic flow of a curve in three space dimensions with a fixed parameter λ exhibits an interesting dynamical behavior. In Figure 4.5 we show projections of the three-dimensional initial curve. The results of the evolution are shown in Figures 4.6–4.9. Because of the high symmetry of the initial curve, the evolution gets near the two-dimensional stable elastica, like the multiply covered circle or the figure eight curve, and stays there for a fairly long time before it continues to unravel.

To finish, let us mention some practical features of our algorithms for the elastic flow of curves. The conditioning of the linear system (4.13), (4.14) should be improved by a diagonal preconditioning process. Instead of the original system for f^{m+1} and

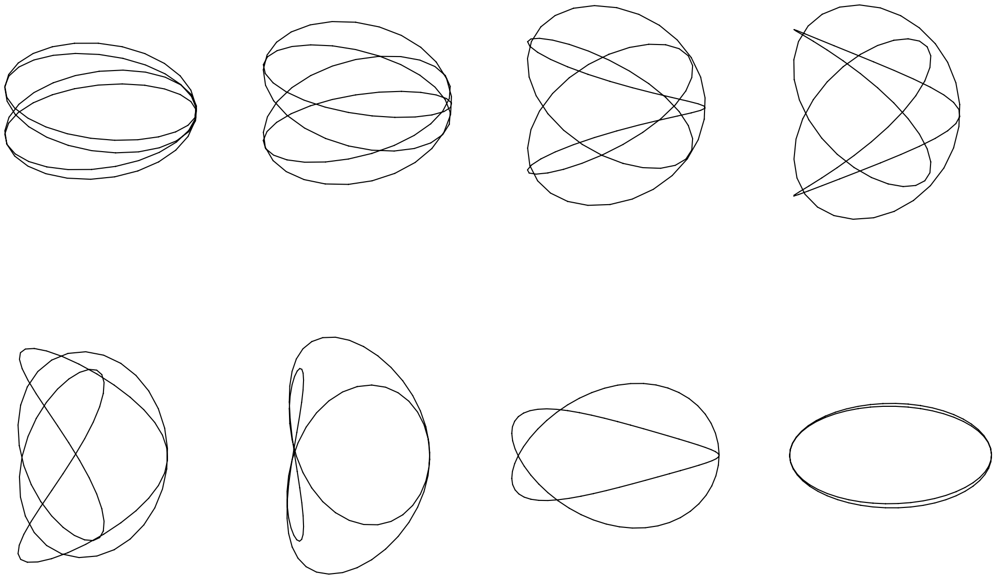


FIG. 4.7. *Continuation of the flow from Figure* 4.6. *From two double circles to one double circle:* $t = 0.611$, 1.01, 1.40, 1.60, 1.79, 1.97, 2.17, 2.37.

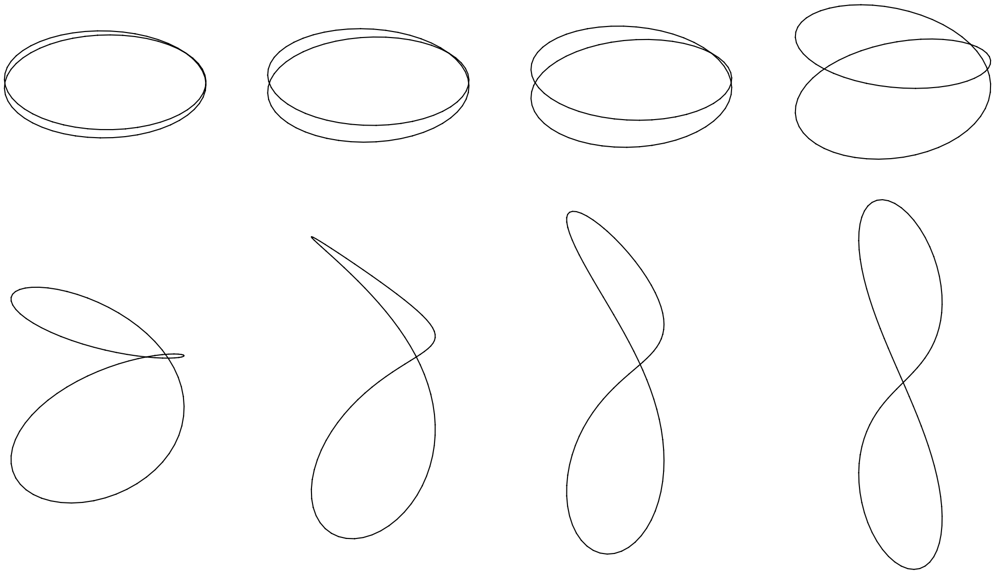


FIG. 4.8. *Continuation of the flow from Figure* 4.7. *From double circle to a figure eight curve:* $t = 3.35$, 3.94, 4.34, 5.13, 5.72, 6.11, 6.31, 6.90.

$\varkappa^{m+1}$, one solves the linear system

$$\left(\frac{1}{\tau}I + D_m^{-\frac{1}{2}}A_m(\beta_m)D_m^{-\frac{1}{2}}\right)\tilde{f}^{m+1} - (D_m^{-\frac{1}{2}}A_mD_m^{-\frac{1}{2}} + \lambda I)\tilde{\varkappa}^{m+1} = \frac{1}{\tau}\tilde{f}^m,$$

$$\tilde{\varkappa}^{m+1} + D_m^{-\frac{1}{2}}A_mD_m^{-\frac{1}{2}}\tilde{f}^{m+1} = 0$$

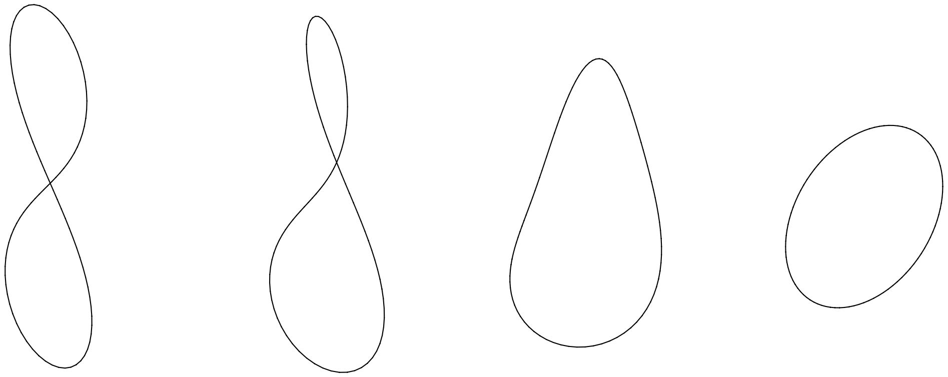


FIG. 4.9. *Continuation of the flow from Figure* 4.8. *From figure eight curve to the final circle:* $t = 11.6,\ 12.4,\ 14.2,\ 16.2$.

for the vectors $\tilde{f}^{m+1} = D_m^{\frac{1}{2}} f^{m+1}$, $\tilde{f}^m = D_m^{\frac{1}{2}} f^m$, and $\tilde{\varkappa}^{m+1} = D_m^{\frac{1}{2}} \varkappa^{m+1}$. For practical purposes it has proved to be of some advantage to redistribute the nodes tangentially along the polygon according to arc length after each time step and thus change the polygon slightly.

REFERENCES

[1] B. Andrews, *The affine curve-lengthening flow*, J. Reine Angew. Math., 506 (1999), pp. 43–83.

[2] T. Aubin, *Nonlinear Analysis on Manifolds. Monge-Ampère Equations*, Grundlehren Math. Wiss. 252, Springer-Verlag, Berlin, Germany, Heidelberg, New York, 1982.

[3] J. W. Cahn and J. E. Taylor, *Surface motion by surface diffusion*, Acta Metallica Materiala, 42 (1994), pp. 1045–1063.

[4] G. Dziuk, *Convergence of a semi–discrete scheme for the curve shortening flow*, Math. Models Methods Appl. Sci., 4 (1994), pp. 589–606.

[5] G. Dziuk, *Discrete anisotropic curve shortening flow*, SIAM J. Numer. Anal., 36 (1999), pp. 1808–1830.

[6] J. Escher, U. F. Mayer, and G. Simonett, *The surface diffusion flow for immersed hypersurfaces*, SIAM J. Math. Anal., 29 (1998), pp. 1419–1433.

[7] N. Koiso, *On the motion of a curve towards elastica*, in Actes de la Table Ronde de Géometrie Différentielle (Luminy 1992), Sémin. Congr. 1, Soc. Math. France, Paris, 1996, pp. 403–436.

[8] J. Langer and D. Singer, *Knotted elastic curves in* $\mathbb{R}^3$, J. London Math. Soc. (2), 30 (1984), pp. 512–520.

[9] J. Langer and D. Singer, *Curve straightening and a minimax argument for closed elastic curves*, Topology, 24 (1985), pp. 75–88.

[10] A. Polden, *Closed Curves of Least Total Curvature*, Preprint 13, SFB 382, Universität Tübingen, Tübingen, Germany, 1995.

[11] A. Polden, *Curves and Surfaces of Least Total Curvature and Fourth-Order Flows*, Ph.D. dissertation, Universität Tübingen, Tübingen, Germany, 1996.

[12] M. Rumpf and A. Schmidt, *GRAPE, Graphics Programming Environment*, Report 8, SFB 256, Universität Bonn, Bonn, Germany, 1990.

[13] C. Truesdell, *The influence of elasticity on analysis: The classical heritage*, Bull. Amer. Math. Soc. (N.S.), 9 (1983), pp. 293–310.

[14] Y. Wen, *Curve straightening flow deforms closed plane curves with nonzero rotation number to circles*, J. Differential Equations, 120 (1995), pp. 89–107.

SIAM J. MATH. ANAL.
Vol. 33, No. 6, pp. 1247–1268

DETECTING AN INCLUSION IN AN ELASTIC BODY BY BOUNDARY MEASUREMENTS*

GIOVANNI ALESSANDRINI[†], ANTONINO MORASSI[‡], AND EDI ROSSET[†]

Abstract. We consider the problem of determining, within an elastic isotropic body Ω, the possible presence of an inclusion D made of different elastic material from boundary measurements of traction and displacement. We prove that the volume of D can be estimated, from above and below, by an easily expressed quantity related to work depending only on the boundary traction and displacement.

Key words. inverse boundary problem, elasticity, size estimates, strong unique continuation

AMS subject classifications. Primary, 35R30; Secondary, 35R25, 73C02

PII. S0036141001388944

1. Introduction.

In this paper we address the following problem of nondestructive testing: *To determine, within an elastic body Ω, the possible presence of an inclusion D made of a different elastic material (i.e., harder or softer) from measurements of traction and displacement taken at the exterior boundary of Ω.*

In mathematical terms, if u denotes the displacement field in Ω, one wishes to recover $D \subset\subset \Omega$ in the *system of linearized elasticity,*

$$\mathrm{div}\,((\chi_{\Omega\setminus D}\mathbb{C} + \chi_D\widetilde{\mathbb{C}})\nabla u) = 0 \quad \text{in } \Omega, \tag{1.1}$$

from the knowledge of one pair of Cauchy data on $\partial\Omega$,

$$(\mathbb{C}\nabla u)\nu = \varphi \quad \text{on } \partial\Omega, \tag{1.2}$$

$$u = g \quad \text{on } \partial\Omega. \tag{1.3}$$

Here $\mathbb{C}$ and $\widetilde{\mathbb{C}}$ denote the elasticity tensor fields in $\Omega \setminus D$ and in D, respectively; ν is the unit exterior normal to $\partial\Omega$; and χ_E denotes the characteristic function of E.

This appears to be an extremely difficult inverse problem. A similar problem in electrical impedance tomography (for which the direct problem involves a single scalar elliptic partial differential equation, rather than a system) has received a great deal of attention in recent years. (See, for instance, Friedman [Fr87], Friedman and Gustafsson [FrG87], Friedman and Isakov [FrI89], Alessandrini, Rosset, and Seo [ARS00] as well as Alessandrini and Isakov [AI96], and Alessandrini [Al99] for an extensive reference list.) Even so, many fundamental questions remain unanswered. One might also consult Ikehata [I98] for previous results on this problem.

Here, following the line of research initiated in Alessandrini and Rosset [AR98], Kang, Seo, and Sheen [KSS97], and Alessandrini, Rosset, and Seo [ARS00] in the electrostatic setting, we pose a relatively modest but realistic goal: *Can we estimate the*

*Received by the editors June 5, 2001; accepted for publication (in revised form) November 19, 2001; published electronically April 26, 2002. Work supported in part by MURST grant MM01111258 and by IAGA-CNR.

http://www.siam.org/journals/sima/33-6/38894.html

[†]Dipartimento di Scienze Matematiche, Università degli Studi di Trieste, Trieste, Italy (alessang@univ.trieste.it, rossedi@univ.trieste.it).

[‡]Dipartimento di Ingegneria Civile, Università degli Studi di Udine, Udine, Italy (antonino.morassi@dic.uniud.it).

size (i.e., volume) of the unknown inclusion D from one set of boundary measurements of traction and displacement?

In the present paper we restrict our attention to the Lamé system of linearized elasticity, corresponding to the system (1.1) when the material is isotropic.

In order to illustrate our main results it is convenient to consider the solution u_0 to the Neumann problem (1.1)–(1.2) when D is the empty set.

Theorem 2.3 below states that if, for a given $h_1 > 0$, the "fatness-condition"

$$| \{x \in D \mid dist(x, \partial D) > h_1\} | \geq \frac{1}{2}|D| \tag{1.4}$$

is satisfied, then

$$C_1 \left| \frac{W - W_0}{W_0} \right| \leqq |D| \leqq C_2 \left| \frac{W - W_0}{W_0} \right|, \tag{1.5}$$

where C_1, C_2 are estimated in terms of the data. Here, the quantities $W = \int_{\partial\Omega} g \cdot \varphi$ and $W_0 = \int_{\partial\Omega} g_0 \cdot \varphi$ represent the work exerted by the surface forces φ when the boundary displacement fields are g and $g_0 = u_0|_{\partial\Omega}$, respectively. See Remark 2.5 for a discussion of the "fatness-condition" (1.4).

In Theorem 2.4 we treat the case when no a priori assumption is made on D. We find that for a suitable $p > 1$, we have

$$C_1 \left| \frac{W - W_0}{W_0} \right| \leqq |D| \leqq C_2 \left| \frac{W - W_0}{W_0} \right|^{\frac{1}{p}}. \tag{1.6}$$

(See section 2 below for the precise statements.)

We believe that these estimates should be useful in practice as a decision tool in quality control tests. Namely, one can fix experimentally a threshold parameter $T > 0$ in such a way to say that D is absent or negligible if $|\frac{W-W_0}{W_0}| \leq T$, whereas D is significant if $|\frac{W-W_0}{W_0}| \geq T$.

The main underlying idea in these estimates is that the integral

$$\int_D |\widehat{\nabla} u_0|^2 \quad \text{is comparable to} \quad |W - W_0|, \tag{1.7}$$

where $\widehat{\nabla} u_0 = \frac{1}{2}(\nabla u_0 + (\nabla u_0)^T)$ is the strain tensor field; see Ikehata [I98].

The next point is to control the above integral in terms of the measure (volume) of D. On the one hand, this task involves upper bounds on $|\widehat{\nabla} u_0|^2$, which is standard in the regularity theory of elliptic systems like (1.1). On the other hand, it involves local lower bounds on $|\widehat{\nabla} u_0|^2$. Rather than regularity theory, this task is more strictly related to the issue of unique continuation, namely, the study of the character of zeros (i.e., order of vanishing and size of the zero sets) of nontrivial solutions to system (1.1). Unique continuation is very well studied and understood for the case of linear elliptic equations. (See, for instance, Aronszajn, Krzywicki, and Szarski [AKS62], Garofalo and Lin [GL86], [GL87], and Koch and Tataru [KT01].) However, until recently, only results of weak unique continuation for the elasticity system were known; see Weck [W69], [W01], Leis [L86], Dehman and Robbiano [DR93], Ang et al. [AITY98], and Eller et al. [EINT00]. In this paper we apply some of the estimates of unique continuation found in [AM01] (three spheres inequalities [AM01, (5.2)] and especially doubling inequalities yielding *strong unique continuation* [AM01, (5.5)]) and we further elaborate on this topic. The main result in this direction here are new *doubling inequalities* for the reference solution u_0 (see Theorem 3.9) and for its symmetrized gradient $\widehat{\nabla} u_0$ (see Corollary 3.10). Such an inequality allows us to prove for $|\widehat{\nabla} u_0|^2$ the property of being a Muckenhoupt weight (Coifman and Fefferman [CF74], Garcia-Cuerva and Rubio de Francia [GCRDF85]). This is a property of homogeneity in the average at all scales which was first proved for solutions of scalar elliptic equations by Garofalo and Lin [GL86].

The plan of the paper is as follows. In section 2 we introduce some notation and state our main results (Theorems 2.3 and 2.4). Section 3 is devoted to the derivation of quantitative estimates of unique continuation for solutions to the Lamé system, following ideas introduced in Alessandrini and Morassi [AM01]. In section 4 we first derive an interior average lower bound on $|\widehat{\nabla} u_0|^2$ on small balls contained inside Ω (see Proposition 4.1). Moreover, we rephrase the doubling inequalities obtained in the previous section in terms of the boundary data (see Proposition 4.3), and we show that $|\widehat{\nabla} u_0|^2$ is a Muckenhoupt weight (see Proposition 4.4). Finally, section 5 contains the proofs of the main theorems.

2. Main results.

Let us introduce some notation which will be useful in what follows. We restrict our attention to the dimensions $n = 2, 3$, which are those physically relevant for elasticity.

Given a bounded domain $\Omega \subset \mathbb{R}^n$, $n = 2, 3$, for any $h > 0$ we shall denote

$$\Omega_h = \{x \in \Omega \mid dist(x, \partial\Omega) > h\}. \tag{2.1}$$

When locally representing a boundary as a graph, it will be convenient to use the following notation. For every $x \in \mathbb{R}^n$ we shall set $x = (x', x_n)$, where $x' \in \mathbb{R}^{n-1}$, $x_n \in \mathbb{R}$.

DEFINITION 2.1. *Given a bounded domain $\Omega \subset \mathbb{R}^n$, we shall say that $\partial\Omega$ is of class $C^{1,1}$ with constants $r_0, M_0 > 0$ if, for any $x_0 \in \partial\Omega$, there exists a rigid transformation of coordinates under which we have $x_0 = 0$ and*

$$\Omega \cap B_{r_0}(0) = \{x \in B_{r_0}(0) \mid x_n > \varphi(x')\},$$

where φ is a $C^{1,1}$ function on $B_{r_0}(0) \subset \mathbb{R}^{n-1}$ satisfying

$$\varphi(0) = |\nabla\varphi(0)| = 0$$

and

$$\|\varphi\|_{C^{1,1}(B_{r_0}(0))} \leq M_0 r_0.$$

Notice that this quantitative formulation of the smoothness of the boundary also involves the introduction of the dimensional parameter r_0, which gives us the scale at which the boundary is representable as a graph.

REMARK 2.1. *We have chosen to normalize all norms in such a way that their terms are dimensionally homogeneous and coincide with the standard definition when the dimensional parameter equals one. For instance, the norm appearing above is meant as follows:*

$$\|\phi\|_{C^{1,1}(B_{r_0}(0))} = \|\phi\|_{L^\infty(B_{r_0}(0))} + r_0\|\nabla\phi\|_{L^\infty(B_{r_0}(0))} + {r_0}^2\|\nabla^2\phi\|_{L^\infty(B_{r_0}(0))}. \tag{2.2}$$

Similarly, given a function $f : \Omega \mapsto \mathbb{R}$, where $\partial\Omega$ satisfies Definition 2.1, we shall denote

$$\|f\|_{C^{1,1}(\Omega)} = \|f\|_{L^\infty(\Omega)} + r_0\|\nabla f\|_{L^\infty(\Omega)} + {r_0}^2\|\nabla^2 f\|_{L^\infty(\Omega)}. \tag{2.3}$$

Notice also that when $\Omega = B_R(0)$, Ω then satisfies Definition 2.1 with $r_0 = R$.

We consider weak solutions $u \in H^1(\Omega, \mathbb{R}^n)$ to the *displacement equation of equilibrium* when body forces are absent:

$$\operatorname{div}(\mathbb{C}(x)(\nabla u(x))) = 0 \quad \text{in } \Omega; \tag{2.4}$$

see Gurtin [Gur72].

We shall assume throughout that the elasticity tensor field $\mathbb{C} = \mathbb{C}(x)$ of the materials under consideration have components C_{ijkl} which satisfy the following conditions:

$$C_{ijkl} \in L^\infty(\Omega, \mathbb{R}) \quad \forall i, j, k, l = 1, \dots, n, \tag{2.5}$$

$$C_{ijkl} = C_{klij} = C_{klji} \quad \forall i, j, k, l = 1, \dots, n, \quad \text{a.e. in} \quad \Omega. \tag{2.6}$$

We recall that the symmetry conditions (2.6) are equivalent to

$$\mathbb{C}A = \mathbb{C}\widehat{A}, \tag{2.7}$$

$$\mathbb{C}A \quad \text{is symmetric}, \tag{2.8}$$

$$\mathbb{C}A \cdot B = \mathbb{C}B \cdot A \tag{2.9}$$

for every pair of $n \times n$ matrices A,B.

Here, and in what follows, the following notation has been used:

$$(\mathbb{C}A)_{ij} = \sum_{k,l=1}^{n} C_{ijkl}A_{kl}, \tag{2.10}$$

$$A \cdot B = \sum_{i,j=1}^{n} A_{ij}B_{ij}, \tag{2.11}$$

$$|A| = (A \cdot A)^{\frac{1}{2}}, \tag{2.12}$$

$$\widehat{A} = \frac{1}{2}(A + A^T), \tag{2.13}$$

for every pair of $n \times n$ matrices A,B.

We shall also use the following conventions for inequalities. Given $\mathbb{C}$, $\widetilde{\mathbb{C}}$ satisfying (2.5), (2.6) we shall say that

$$\widetilde{\mathbb{C}} \leq \mathbb{C} \tag{2.14}$$

if and only if

$$\widetilde{\mathbb{C}}A \cdot A \leq \mathbb{C}A \cdot A \tag{2.15}$$

for every *symmetric* $n \times n$ matrix A.

We shall say that $\mathbb{C}$ is *strongly convex* in Ω if there exists a positive constant ξ_0 such that

$$\mathbb{C}(x)A \cdot A \geq \xi_0 |A|^2 \quad \text{for a.e.} \quad x \in \Omega, \tag{2.16}$$

for any symmetric $n \times n$ matrix A.

$\mathbb{C}$ is said to be *strongly elliptic* in Ω if there exists a positive constant κ_0 such that

$$\mathbb{C}(x)A \cdot A \geq \kappa_0 |A|^2 \quad \text{for a.e.} \quad x \in \Omega, \tag{2.17}$$

for any matrix A of the form $A_{ij} = a_i b_j$, where a and b are n-vectors. It is well known that *if* $\mathbb{C}$ *is strongly convex, then it is also strongly elliptic.*

When the elastic material is *isotropic*, then the elasticity tensor $\mathbb{C}$ takes the following form:

$$C_{ijkl}(x) = \lambda(x)\delta_{ij}\delta_{kl} + \mu(x)(\delta_{ki}\delta_{lj} + \delta_{li}\delta_{kj}), \tag{2.18}$$

where $\lambda = \lambda(x)$ and $\mu = \mu(x) \in L^\infty(\Omega, \mathbb{R})$ are the *Lamé moduli.* Hence, in this case, denoting by I_n the $n \times n$ identity matrix, we have

$$\mathbb{C}(x)A = \lambda(x)(A \cdot I_n)I_n + 2\mu(x)\widehat{A}, \tag{2.19}$$

and the displacement equation of equilibrium (2.4) becomes the Lamé system

$$\mathrm{div}\,(2\mu\widehat{\nabla}u) + \nabla(\lambda \mathrm{div}\, u) = 0 \quad \text{in } \Omega. \tag{2.20}$$

In the isotropic case, the *strong convexity* condition takes the form

$$\mu(x) \geq \alpha_0, \quad 2\mu(x) + n\lambda(x) \geq \gamma_0 \quad \text{for a.e.} \quad x \in \Omega \tag{2.21}$$

and the *strong ellipticity* condition is expressed by

$$\mu(x) \geq \alpha_0, \quad 2\mu(x) + \lambda(x) \geq \beta_0 \quad \text{for a.e.} \quad x \in \Omega, \tag{2.22}$$

where α_0, β_0, γ_0 are positive constants.

Let Ω be a bounded domain whose boundary is of class $C^{1,1}$ with given constants $r_0, M_0 > 0$. Let D be a measurable, possibly disconnected, subset of Ω such that, given $d_0 > 0$,

$$dist(D, \partial\Omega) \geq d_0. \tag{2.23}$$

Given elasticity tensors $\mathbb{C}$, $\widetilde{\mathbb{C}}$ satisfying (2.5), (2.6) we shall consider traction problems in Ω when the elasticity tensor is either $\chi_{\Omega\setminus D}\mathbb{C} + \chi_D\widetilde{\mathbb{C}}$ or $\mathbb{C}$.

We shall prescribe a boundary traction field $\varphi \in L^2(\partial\Omega, \mathbb{R}^n)$ satisfying the compatibility conditions

$$\int_{\partial\Omega} \varphi \cdot r = 0 \tag{2.24}$$

for every *infinitesimal rigid displacement* r, that is, $r(x) = c + Wx$, where c is any constant n-vector and W is any constant skew $n \times n$ matrix. Namely we shall consider weak solutions $u, u_0 \in H^1(\Omega, \mathbb{R}^n)$ of the following problems:

$$\operatorname{div}\left((\chi_{\Omega \setminus D}\mathbb{C} + \chi_D \widetilde{\mathbb{C}})\nabla u\right) = 0 \quad \text{in } \Omega, \tag{2.25}$$

$$(\mathbb{C}\nabla u)\nu = \varphi \quad \text{on } \partial\Omega; \tag{2.26}$$

$$\operatorname{div}(\mathbb{C}\nabla u_0) = 0 \quad \text{in } \Omega, \tag{2.27}$$

$$(\mathbb{C}\nabla u_0)\nu = \varphi \quad \text{on } \partial\Omega. \tag{2.28}$$

Regarding existence, we recall that, provided the compatibility condition (2.24) is satisfied, a solution of the traction problem exists as long as the involved elasticity tensor either satisfies the strong convexity condition or is continuous and satisfies the strong ellipticity condition; see, for instance, Valent [V88, section III].

With respect to uniqueness we recall that it is well known that solutions u, u_0 to the above problems are uniquely determined up to an infinitesimal rigid displacement. In order to uniquely identify such solutions, we shall assume from now on that both u and u_0 satisfy the following normalization conditions:

$$\int_\Omega u = 0, \quad \int_\Omega (\nabla u - (\nabla u)^T) = 0. \tag{2.29}$$

We set g, $g_0 \in H^{1/2}(\partial\Omega, \mathbb{R}^n)$ to be the traces of u, u_0, respectively, on $\partial\Omega$.

Now we are in position to state our main result on the estimates for the size of the inclusion.

We shall use the following assumptions on the elasticity tensors $\mathbb{C}$, $\widetilde{\mathbb{C}}$:

(i) $\mathbb{C}$ satisfies the isotropy condition (2.18) and the strong convexity (2.21);

(ii) (bounds on the jump and uniform strong convexity for $\widetilde{\mathbb{C}}$)
either there exist $\eta > 0$ and $\delta > 1$ such that

$$\eta\mathbb{C} \le \widetilde{\mathbb{C}} - \mathbb{C} \le (\delta - 1)\mathbb{C} \quad \text{a.e. in} \quad \Omega \tag{2.30}$$

or there exist $\eta > 0$ and $0 < \delta < 1$ such that

$$-(1-\delta)\mathbb{C} \le \widetilde{\mathbb{C}} - \mathbb{C} \le -\eta\mathbb{C} \quad \text{a.e. in} \quad \Omega; \tag{2.31}$$

(iii) ($C^{1,1}$ regularity for $\mathbb{C}$)
there exists $M > 0$ such that

$$\|\mu\|_{C^{1,1}(\Omega)} + \|\lambda\|_{C^{1,1}(\Omega)} \le M. \tag{2.32}$$

Remark 2.2. *It is worth noticing that very mild assumptions are made on the unknown inclusion; namely, the inclusion D may consist of an anisotropic material which is either harder (case* (2.30)*) or softer (case* (2.31)*) than the surrounding material in Ω, and no additional regularity assumption is required on the elasticity tensor inside D.*

THEOREM 2.3. *Let Ω be a bounded domain in $\mathbb{R}^n$ such that $\partial\Omega$ is of class $C^{1,1}$ with constants r_0, M_0. Let D be a measurable subset of Ω satisfying* (2.23) *and*

$$|D_{h_1}| \geq \frac{1}{2}\,|D| \tag{2.33}$$

for a given positive constant h_1. Let $\mathbb{C}$, $\widetilde{\mathbb{C}}$ satisfy (i), (ii), (iii). *If* (2.30) *holds, then we have*

$$\frac{1}{\delta-1}C_1^+\frac{\int_{\partial\Omega}(g_0-g)\cdot\varphi}{\int_{\partial\Omega}g_0\cdot\varphi} \leq |D| \leq \frac{\delta}{\eta}C_2^+\frac{\int_{\partial\Omega}(g_0-g)\cdot\varphi}{\int_{\partial\Omega}g_0\cdot\varphi}. \tag{2.34}$$

If, conversely, (2.31) *holds, then we have*

$$\frac{\delta}{1-\delta}C_1^-\frac{\int_{\partial\Omega}(g-g_0)\cdot\varphi}{\int_{\partial\Omega}g_0\cdot\varphi} \leq |D| \leq \frac{1}{\eta}C_2^-\frac{\int_{\partial\Omega}(g-g_0)\cdot\varphi}{\int_{\partial\Omega}g_0\cdot\varphi}, \tag{2.35}$$

where C_1^+, C_1^- depend only on the geometrical parameters $|\Omega|$, r_0, M_0 (see Definition 2.1*), d_0 (see* (2.23)*) and on the bounds on the Lamé moduli α_0, γ_0 (see* (2.21)*), M (see* (2.32)*), and C_2^+, C_2^- depend only on the same quantities and also on h_1 and on the ratio $\|\varphi\|_{L^2(\partial\Omega)}/\|\varphi\|_{H^{-1/2}(\partial\Omega)}$.*

THEOREM 2.4. *Let Ω be as in Theorem* 2.3 *and let D be any measurable subset of Ω satisfying* (2.23). *Let $\mathbb{C}$, $\widetilde{\mathbb{C}}$ satisfy* (i), (ii), (iii). *If* (2.30) *holds, then we have*

$$\frac{1}{\delta-1}C_1^+\frac{\int_{\partial\Omega}(g_0-g)\cdot\varphi}{\int_{\partial\Omega}g_0\cdot\varphi} \leq |D| \leq \left(\frac{\delta}{\eta}\right)^{\frac{1}{p}}C_2^+\left(\frac{\int_{\partial\Omega}(g_0-g)\cdot\varphi}{\int_{\partial\Omega}g_0\cdot\varphi}\right)^{\frac{1}{p}}. \tag{2.36}$$

If, conversely, (2.31) *holds, then we have*

$$\frac{\delta}{1-\delta}C_1^-\frac{\int_{\partial\Omega}(g-g_0)\cdot\varphi}{\int_{\partial\Omega}g_0\cdot\varphi} \leq |D| \leq \left(\frac{1}{\eta}\right)^{\frac{1}{p}}C_2^-\left(\frac{\int_{\partial\Omega}(g-g_0)\cdot\varphi}{\int_{\partial\Omega}g_0\cdot\varphi}\right)^{\frac{1}{p}}, \tag{2.37}$$

where C_1^+, C_1^- are the same as in Theorem 2.3*, whereas $p>1$, C_2^+, C_2^- depend only on $|\Omega|$, r_0, M_0, d_0, α_0, γ_0, M, and $\|\varphi\|_{L^2(\partial\Omega)}/\|\varphi\|_{H^{-1/2}(\partial\Omega)}$.*

REMARK 2.5. *Let us notice that the "fatness-condition"* (2.33) *is satisfied when mild a priori regularity assumptions are made on D. For instance, the constant h_1 can be easily estimated when D is a priori known to be Lipschitz; we refer to Alessandrini and Rosset* [AR98, Lemma 2.8] *for related calculations. See also Alessandrini, Rosset, and Seo* [ARS00] *for comments on the optimality of this kind of results in the case of a scalar elliptic equation.*

3. Quantitative estimates of unique continuation. In this section we shall prove quantitative estimates of unique continuation in the form of *three spheres inequalities* and *doubling inequalities* for solutions $u \in H^1(\Omega, \mathbb{R}^n)$ to the Lamé system of linearized elasticity (2.20) in a bounded domain Ω satisfying Definition 2.1 with constants r_0, M_0. Throughout this section the Lamé moduli $\mu = \mu(x)$, $\lambda = \lambda(x)$ are assumed to satisfy the strong ellipticity condition (2.22) and the regularity assumption (2.32). Following ideas introduced in [AM01], the first step consists of reducing the Lamé system (2.20) to a weakly coupled elliptic system with Laplacian principal part. We denote by $\mathbb{M}^{m\times n}$ the space of $m \times n$ real valued matrices.

PROPOSITION 3.1. *Under the above assumptions, there exist $B \in L^\infty(\Omega, L(\mathbb{M}^{(n+1)\times n}, \mathbb{R}^{n+1}))$ and $V \in L^\infty(\Omega, L(\mathbb{R}^{n+1}, \mathbb{R}^{n+1}))$ such that, for every weak solution $u \in H^1(\Omega, \mathbb{R}^n)$ to* (2.20)*, the $\mathbb{R}^{n+1}$-valued function U given by*

$$U = \begin{pmatrix} u \\ r_0 \operatorname{div} u \end{pmatrix} \tag{3.1}$$

belongs to $W^{2,p}_{loc}(\Omega, \mathbb{R}^{n+1})$ for every $p < \infty$ and satisfies

$$-\Delta U + B(\nabla U) + V(U) = 0 \quad \text{in } \Omega. \tag{3.2}$$

Moreover

$$r_0 \|B\|_{L^\infty(\Omega, L(\mathbb{M}^{(n+1)\times n}, \mathbb{R}^{n+1}))} + r_0^2 \|V\|_{L^\infty(\Omega, L(\mathbb{R}^{n+1}, \mathbb{R}^{n+1}))} \leqslant CM, \tag{3.3}$$

where $C > 0$ depends only on α_0 and β_0.

Proof. The proof is essentially contained in [AM01, Theorem 2.1]. Here the statement is slightly modified in order to encompass the scaling invariance of the norms introduced in the present paper. □

Three spheres inequalities and doubling inequalities for solutions u to systems of the form (3.2), under the assumption (3.3), were derived in [AM01, Theorems 3.1 and 4.1]. Next, one can obtain analogous estimates for solutions u to the Lamé system (2.20) via the reduction described in Proposition 3.1.

PROPOSITION 3.2 (see [AM01, Theorem 5.1]). *Let $\Omega = B_R = \{x \in \mathbb{R}^n \mid |x| < R\}$. Under the above assumptions, there exists $\overline{\theta}$, $0 < \overline{\theta} \leqslant 1$, depending only on α_0, β_0, M, such that for every weak solution $u \in H^1(B_R, \mathbb{R}^n)$ to* (2.20) *and for every r_1, r_2, r_3, $0 < r_1 < r_2 < r_3 \leqslant \overline{\theta} R$, we have*

$$\int_{B_{r_2}} |u|^2 \leqslant C \left(\int_{B_{r_1}} |u|^2 \right)^\delta \left(\int_{B_{r_3}} |u|^2 \right)^{1-\delta}, \tag{3.4}$$

where $C > 0$, $0 < \delta < 1$, depend only on α_0, β_0, M, $\frac{r_1}{r_3}$, and $\frac{r_2}{r_3}$.

Proof. The proof can be found in [AM01]. We notice that here, in view of our scaling on the norms (see Remark 2.1), the constant C does not explicitly depend on R. □

In view of the applications in section 5, we need the analogous result for $\widehat{\nabla} u$.

COROLLARY 3.3. *Under the same hypotheses of Proposition* 3.2, *for every weak solution $u \in H^1(B_R, \mathbb{R}^n)$ to* (2.20) *and for every r_1, r_2, r_3, $0 < r_1 < r_2 < r_3 \leqslant \overline{\theta} R$, we have*

$$\int_{B_{r_2}} |\widehat{\nabla} u|^2 \leqslant C \left(\int_{B_{r_1}} |\widehat{\nabla} u|^2 \right)^\delta \left(\int_{B_{r_3}} |\widehat{\nabla} u|^2 \right)^{1-\delta}, \tag{3.5}$$

where $\overline{\theta}$, $0 < \overline{\theta} \leqslant 1$, is the same as in Proposition 3.2 *and $C > 0$, $0 < \delta < 1$ depend only on α_0, β_0, M, $\frac{r_1}{r_3}$, and $\frac{r_2}{r_3}$.*

In order to prove Corollary 3.3 it is convenient to recall the following two inequalities.

LEMMA 3.4 (Caccioppoli-type inequality). *If $\mathbb{C}$ satisfies* (2.18), (2.22), *and* (2.32), *then for every solution $u \in H^1(B_R, \mathbb{R}^n)$ to* (2.4) *and for every r, $0 < r < R$,* *we have*

$$\int_{B_r} |\nabla u|^2 \leqslant \frac{C}{(R-r)^2} \int_{B_R} |u|^2, \tag{3.6}$$

where $C > 0$ *depends only on* α_0, β_0, M.

Proof. The proof follows by a standard cut-off argument from Gärding's inequality [V88]. □

Given $u \in H^1(B_R, \mathbb{R}^n)$ and r, $0 < r < R$, set

$$u_r = \frac{1}{|B_r|} \int_{B_r} u, \tag{3.7}$$

$$W_r = \frac{1}{2|B_r|} \int_{B_r} (\nabla u - (\nabla u)^T). \tag{3.8}$$

LEMMA 3.5 (Korn inequality). *There exists an absolute constant* $C > 0$ *such that for every* $u \in H^1(B_R, \mathbb{R}^n)$ *and every* r, $0 < r < R$, *we have*

$$\int_{B_R} |\nabla u - W_r|^2 + \frac{1}{R^2}|u - u_r - W_r x|^2 \leqslant C \left(\frac{R}{r}\right)^{4n-2} \int_{B_R} |\widehat{\nabla} u|^2. \tag{3.9}$$

REMARK 3.6. *When* $r = R$ *this is the well-known second Korn inequality, which is known to hold in every sufficiently regular domain* Ω *(see* [Fi72], [T99]*). Here we introduce a minor variant, in which the averages of* u *and of the skew part of* ∇u *are taken on the smaller ball* B_r*. For the convenience of the reader, a sketch of the main arguments of a proof is outlined at the end of this section.*

Proof of Corollary 3.3. The function v defined in B_R by

$$v = u - u_{r_1} - W_{r_1} x \tag{3.10}$$

satisfies (2.20) and $\frac{1}{|B_{r_1}|}\int_{B_{r_1}} v = 0$, $\frac{1}{2|B_{r_1}|}\int_{B_{r_1}} \nabla v - (\nabla v)^T = 0$. By applying to v the Caccioppoli-type inequality (3.6) and the three spheres inequality (3.4) and using the Korn inequality (3.9) twice, we have

$$\begin{aligned} \int_{B_{r_2}} |\widehat{\nabla} u|^2 = \int_{B_{r_2}} |\widehat{\nabla} v|^2 &\leqslant \frac{C_1}{(r_3 - r_2)^2} \int_{B_{\frac{r_2+r_3}{2}}} |v|^2 \\ &\leqslant \frac{C_2}{(r_3 - r_2)^2} \left(\int_{B_{r_1}} |v|^2\right)^{\delta} \left(\int_{B_{r_3}} |v|^2\right)^{1-\delta} \\ &\leqslant C_3 \left(\int_{B_{r_1}} |\widehat{\nabla} u|^2\right)^{\delta} \left(\int_{B_{r_3}} |\widehat{\nabla} u|^2\right)^{1-\delta}, \end{aligned} \tag{3.11}$$

where C_1, C_2, C_3 are constants depending only on α_0, β_0, M, $\frac{r_1}{r_3}$, and $\frac{r_2}{r_3}$. □

In order to obtain the doubling inequality for solutions to the Lamé system (2.20), we need to state a slightly modified version of the doubling inequality for solutions U to (3.2) contained in [AM01, Theorem 4.1].

Proposition 3.7. *Let $\Omega = B_R$ and let $R\|B\|_\infty + R^2\|V\|_\infty \leqslant E$. There exists θ^*, $0 < \theta^* \leqslant 1$, depending only on E, such that for every nonzero solution $U \in H^1(B_R, \mathbb{R}^{n+1})$ to (3.2) we have*

$$\int_{B_{2r}} |U|^2 \leqslant C \int_{B_r} |U|^2 \text{ for every } r,\ 0 < r \leqslant \frac{\theta^* R}{2}, \tag{3.12}$$

where C depends only on E and $N_0(\theta^* R)$, where

$$N_0(r) = N_0(U; r) = \frac{r\int_{B_r} |\nabla U|^2}{\int_{\partial B_r} |U|^2}, \qquad 0 < r \leqslant R. \tag{3.13}$$

Moreover C is increasing with $N_0(\theta^* R)$.

Proof. By Theorem 4.1 in [AM01] and by a rescaling argument, it easily follows that (3.12) holds, with C depending only on E and on $N(\theta^* R)$, where

$$N(r) = N(U; r) = \frac{r\int_{B_r} |\nabla U|^2 + B(\nabla U)\cdot U + V(U)\cdot U}{\int_{\partial B_r} |U|^2}, \qquad 0 < r \leqslant R, \tag{3.14}$$

the dependence on this last variable being monotonically increasing. Hence, we have to show that $N(r)$ can be bounded from above in terms of $N_0(r)$. It is convenient to recall the following notation introduced in [AM01]:

$$G(r) = \int_{B_r} |U|^2, \tag{3.15}$$

$$H(r) = \int_{\partial B_r} |U|^2, \tag{3.16}$$

$$I(r) = \int_{B_r} |\nabla U|^2 + B(\nabla U)\cdot U + V(U)\cdot U, \tag{3.17}$$

$$D(r) = \int_{B_r} |\nabla U|^2 \tag{3.18}$$

for $0 < r \leqslant R$. We easily have

$$I(r) \leqslant D(r) + \frac{E}{R}(D(r)G(r))^{\frac{1}{2}} + \frac{E}{R^2}G(r) \leqslant C\left(D(r) + \frac{G(r)}{R^2}\right), \tag{3.19}$$

with C depending only on E. Moreover, from Lemma 3.3 in [AM01] we have

$$G(r) \leqslant rH(r) \text{ for } r \leqslant \theta^* R. \tag{3.20}$$

Hence, for $r \leqslant \theta^* R$ we have

$$N(r) = \frac{rI(r)}{H(r)} \leqslant C\left(N_0(r) + \frac{r^2}{R^2}\right) \leqslant C(N_0(r) + 1), \tag{3.21}$$

where C depends only on E. □

REMARK 3.8. *Let us notice that analogously to* (3.21) *we also have*

$$N_0(r) \leqslant C(N(r)+1), \text{ for } r \leqslant \theta^* R, \tag{3.22}$$

where C depends only on E.

THEOREM 3.9. *Let $\Omega = B_R$. There exists θ^*, $0 < \theta^* \leqslant 1$, depending only on α_0, β_0, M, such that for every nonzero weak solution $u \in H^1(B_R, \mathbb{R}^n)$ to* (2.20) *we have*

$$\int_{B_{2r}} |u|^2 \leqslant K \int_{B_r} |u|^2 \text{ for every } r,\ 0 < r \leqslant \frac{\theta^* R}{2}, \tag{3.23}$$

where $K > 0$ depends only on α_0, β_0, M and on $\tilde{N}_0(\theta^* R)$, where

$$\tilde{N}_0(r) = \frac{r^2 \int_{B_r} |\nabla u|^2 + R^2 |\nabla(\mathrm{div} u)|^2}{\int_{B_r} |u|^2 + R^2 |\mathrm{div} u|^2}, \qquad 0 < r \leqslant R. \tag{3.24}$$

Moreover K is increasing with $\tilde{N}_0(\theta^* R)$.

Proof. By applying (3.12) of Proposition 3.7 to the solution U to (3.2) given by the position (3.1), and recalling (3.3), we have

$$\int_{B_{2r}} |u|^2 + R^2 |\mathrm{div}\, u|^2 \leqslant C \int_{B_r} |u|^2 + R^2 |\mathrm{div}\, u|^2 \text{ for every } r,\ 0 < r \leqslant \frac{\theta^* R}{2}, \tag{3.25}$$

where C depends only on α_0, β_0, M and on $N_0(\theta^* R)$, with $N_0(r)$ given by (3.13), the dependence on this last variable being monotonically increasing. By an iterated application of (3.25) and by the Caccioppoli-type inequality (3.6), we have

$$\int_{B_{2r}} |u|^2 \leqslant C\left(1 + \frac{R^2}{r^2}\right) \int_{B_r} |u|^2 \text{ for every } r,\ 0 < r \leqslant \frac{\theta^* R}{2}. \tag{3.26}$$

Let ρ be such that $0 < \rho \leqslant 1$, and let

$$u_\rho(x) = u(\rho x) \text{ in } B_{\frac{R}{\rho}}.$$

We have that u_ρ is a solution in $B_{\frac{R}{\rho}}$ to the Lamé system (2.20) with Lamé moduli satisfying uniformly the bounds (2.22), (2.32). Therefore, by (3.26) we have that

$$\int_{B_{2r}} |u_\rho|^2 \leqslant C_\rho \left(1 + \frac{R^2}{\rho^2 r^2}\right) \int_{B_r} |u_\rho|^2 \text{ for every } r,\ 0 < r \leqslant \frac{\theta^* R}{2\rho}, \tag{3.27}$$

where C_ρ depends only on α_0, β_0, M, and $N_0(U_\rho; \frac{\theta^* R}{\rho})$, where U_ρ is given by (3.1) when u, r_0 are replaced with u_ρ, $\frac{R}{\rho}$, respectively. Here, again, C_ρ is increasing with $N_0(U_\rho; \frac{\theta^* R}{\rho})$. We have for any r, $0 < r \leqslant \frac{\theta^* R}{2\rho}$,

$$\begin{aligned} N_0(U_\rho; r) &= \frac{r \int_{B_r} |\nabla u_\rho|^2 + \frac{R^2}{\rho^2} |\nabla \mathrm{div}\, u_\rho|^2}{\int_{\partial B_r} |u_\rho|^2 + \frac{R^2}{\rho^2} |\mathrm{div}\, u_\rho|^2} \\ &= \frac{r \int_{B_r} \rho^2 |\nabla u(\rho x)|^2 + \frac{R^2}{\rho^2} \cdot \rho^4 |\nabla \mathrm{div}\, u(\rho x)|^2}{\int_{\partial B_r} |u(\rho x)|^2 + \frac{R^2}{\rho^2} \cdot \rho^2 |\mathrm{div}\, u(\rho x)|^2} \\ &= \frac{\rho r \int_{B_{\rho r}} |\nabla u|^2 + R^2 |\nabla \mathrm{div}\, u|^2}{\int_{\partial B_{\rho r}} |u|^2 + R^2 |\mathrm{div}\, u|^2} = N_0(U; \rho r). \end{aligned} \tag{3.28}$$

Hence, in particular,

$$N_0\left(U_\rho;\frac{\theta^* R}{\rho}\right)=N_0(U;\theta^* R). \tag{3.29}$$

Consequently, the quantity C_ρ appearing in (3.27) is uniformly bounded from above with respect to $\rho\in(0,1]$. Taking $r=\frac{\theta^* R}{2\rho}$ in (3.27) and setting $s=r\rho$, we obtain

$$\int_{B_{2s}}|u|^2\leqslant K\int_{B_s}|u|^2 \text{ for every } s,\ 0<s\leqslant\frac{\theta^* R}{2}, \tag{3.30}$$

where K depends only on α_0,β_0,M and on $N_0(U;\theta^* R)$, the dependence on this last variable being monotonically increasing. Recalling (3.13), (3.20), (3.24), we have that $N_0(U;\theta^* R)\leqslant\tilde{N}_0(\theta^* R)$. □

From Theorem 3.9, by using arguments analogous to those employed in the proof of Corollary 3.3, the following doubling inequality for $\widehat{\nabla}u$ follows.

Corollary 3.10. *Let $\Omega=B_R$. There exists θ^*, $0<\theta^*\leqslant 1$, depending only on α_0,β_0,M, such that for every nonzero weak solution $u\in H^1(B_R,\mathbb{R}^n)$ to* (2.20) *we have*

$$\int_{B_{2r}}|\widehat{\nabla}u|^2\leqslant K_r\int_{B_r}|\widehat{\nabla}u|^2 \textit{ for every } r,\ 0<r\leqslant\frac{\theta^* R}{4}, \tag{3.31}$$

where $K_r>0$ depends only on α_0,β_0,M, and $\tilde{N}_0(v;\theta^* R)$ given by (3.24), *where $v=u-u_r-W_r x$, with u_r and W_r defined by* (3.7) *and* (3.8), *respectively. Moreover K_r is increasing with $\tilde{N}_0(v;\theta^* R)$.*

Proof of Lemma 3.5. We adapt arguments from Tiero [T99]. Inequality (3.9) follows, through the introduction of the axial vector ω associated with the skew matrix $W=\frac{\nabla u-(\nabla u)^T}{2}$, from the two scalar inequalities

$$\int_{B_R}(\psi-\psi_r)^2\leqslant C\left(\frac{R}{r}\right)^{2(n-1)}R^2\int_{B_R}|\nabla\psi|^2 \text{ for every } \psi\in H^1(B_R), \tag{3.32}$$

$$\int_{B_R}(\psi-\psi_r)^2\leqslant C\left(\frac{R}{r}\right)^{2n}\|\nabla\psi\|^2_{H^{-1}(B_R)} \text{ for every } \psi\in L^2(B_R). \tag{3.33}$$

Here $C>0$ is an absolute constant and the $H^{-1}(B_R)$-norm above is defined as follows:

$$\|F\|_{H^{-1}(B_R)}=\sup\left\{\int_{B_R}FG\mid G\in H^1(B_R,\mathbb{R}^n),\int_{B_R}|\nabla G|^2=1\right\}.$$

It suffices to prove (3.32), (3.33) when $R=1$ and $\psi\in C^1(\overline{B}_1)$ by usual scaling and density arguments. We recall that (3.32), (3.33) are well known when $r=1$; see, for instance, [MS58]. Let us estimate $\psi_1-\psi_r$ for $0<r<1$. We easily obtain

$$\psi_1-\psi_r=\frac{1}{n\omega_n}\int_r^1 ds\int_{B_1}\nabla\psi(sx)\cdot x dx,$$

and then, by changing variables and reversing the order of integration, we find

$$\psi_1-\psi_r=\frac{1}{n\omega_n}\int_{B_1}\nabla\psi\cdot z,$$

where

$$z(x) = \frac{1}{n}\{\max\{r, |x|\}^{-n} - 1\}x.$$

We have

$$\int_{B_1} |z|^2 \leqslant Cr^{1-n},$$

$$\int_{B_1} |\nabla z|^2 \leqslant Cr^{-n}.$$

Hence

$$|\psi_1 - \psi_r|^2 \leqslant Cr^{2-2n} \int_{B_1} |\nabla \psi|^2,$$

$$|\psi_1 - \psi_r|^2 \leqslant Cr^{-2n} \|\nabla \psi\|^2_{H^{-1}(B_1)},$$

and (3.32), (3.33) follow. □

4. Estimates in terms of the boundary data. In this section we shall consider the traction problem (2.27), (2.28) for a given $\varphi \in L^2(\partial\Omega, \mathbb{R}^n)$ satisfying (2.24). For simplicity of notation we shall denote by u the solution (instead of u_0). The normalization (2.29) is understood throughout.

Regarding the elasticity tensor $\mathbb{C}$ we assume the isotropy condition (2.18), the strong ellipticity (2.22), and the $C^{1,1}$ regularity (2.32).

PROPOSITION 4.1 (Lipschitz propagation of smallness). *For every $\rho > 0$ and for every $x \in \Omega_{\frac{4\rho}{\overline{\theta}}}$, we have*

$$\int_{B_\rho(x)} |\widehat{\nabla} u|^2 \geqslant C_\rho \int_\Omega |\widehat{\nabla} u|^2, \tag{4.1}$$

where $\overline{\theta}$, $0 < \overline{\theta} < 1$, depends only on α_0, β_0, M and C_ρ depends only on α_0, β_0, M, $|\Omega|$, r_0, M_0, $\|\varphi\|_{L^2(\partial\Omega)}/\|\varphi\|_{H^{-1/2}(\partial\Omega)}$, and ρ.

We adapt arguments from [ARS00, Theorem 2.2]. We start with the following auxiliary lemma.

LEMMA 4.2.

$$\int_{\Omega\setminus\Omega_{\frac{5\rho}{\overline{\theta}}}} |\widehat{\nabla} u|^2 \leqslant C\rho^{1/n} \|\varphi\|^2_{L^2(\partial\Omega)}, \tag{4.2}$$

where C depends only on $\alpha_0, \beta_0, M, r_0, M_0, |\Omega|$.

Proof of Lemma 4.2. By Hölder's inequality

$$\|\widehat{\nabla} u\|^2_{L^2(\Omega\setminus\Omega_{\frac{5\rho}{\overline{\theta}}})} \leqslant |\Omega \setminus \Omega_{\frac{5\rho}{\overline{\theta}}}|^{1/n} \|\widehat{\nabla} u\|^2_{L^{2n/(n-1)}(\Omega\setminus\Omega_{\frac{5\rho}{\overline{\theta}}})}, \tag{4.3}$$

and by the Sobolev inequality (see, for instance, [Ad75])

$$\|\widehat{\nabla} u\|^2_{L^{2n/(n-1)}(\Omega)} \leqslant C\|\widehat{\nabla} u\|^2_{H^{1/2}(\Omega)}, \tag{4.4}$$

we have

$$\|\widehat{\nabla} u\|^2_{L^2(\Omega\setminus\Omega_{\frac{5\rho}{\theta}})} \leqslant C|\Omega\setminus\Omega_{\frac{5\rho}{\theta}}|^{1/n}\|u\|^2_{H^{3/2}(\Omega)}, \tag{4.5}$$

where C depends only on r_0, M_0, $|\Omega|$.

Moreover, we have

$$\|u\|_{H^{3/2}(\Omega)} \leqslant C\|\varphi\|_{L^2(\partial\Omega)}, \tag{4.6}$$

where C depends only on $\alpha_0, \beta_0, M, r_0, M_0, |\Omega|$. Inequality (4.6) follows, by interpolation (see [LM72]), from the global estimates for the Neumann problem

$$\|u\|_{H^1(\Omega)} \leqslant C_1\|\varphi\|_{H^{-1/2}(\partial\Omega)}, \tag{4.7}$$

$$\|u\|_{H^2(\Omega)} \leqslant C_2\|\varphi\|_{H^{1/2}(\partial\Omega)}, \tag{4.8}$$

where C_1 and C_2 depend only on α_0, β_0, M, r_0, M_0, $|\Omega|$ (see [ADN64]).

Moreover,

$$|\Omega\setminus\Omega_{\frac{5\rho}{\theta}}| \leqslant C\rho, \tag{4.9}$$

where C depends only on r_0, M_0, $|\Omega|$. (See (A.3) in [AR98] for details.) From (4.5), (4.6), and (4.9) the thesis follows. □

Proof of Proposition 4.1. Let us fix ρ_0, depending only on r_0, M_0, such that $\Omega_{\frac{4\rho}{\theta}}$ is connected for every $\rho \leqslant \rho_0$. Without loss of generality, we may assume $\rho \leqslant \rho_0$ for this proof. Given any $y \in \Omega_{\frac{4\rho}{\theta}}$, let γ be an arc in $\Omega_{\frac{4\rho}{\theta}}$ joining x and y. Let us define $\{x_i\}$, $i = 1, \dots, L$, as follows: $x_1 = x$, $x_{i+1} = \gamma(t_i)$, where $t_i = \max\{t \mid |\gamma(t) - x_i| = 2\rho\}$ if $|x_i - y| > 2\rho$; otherwise let $i = L$ and stop the process. Then, by construction, the balls $B_\rho(x_i)$ are pairwise disjoint, $|x_{i+1} - x_i| = 2\rho$ for $i = 1, \dots, L-1$, $|x_L - y| \leqslant 2\rho$.

Since $x_i \in \Omega_{\frac{4\rho}{\theta}}$, we may apply (3.5) for $x = x_i$, $r_1 = \rho$, $r_2 = 3\rho$, $r_3 = 4\rho$, for $i = 1, \dots, L-1$, obtaining

$$\frac{\|\widehat{\nabla} u\|_{L^2(B_\rho(x_{i+1}))}}{\|\widehat{\nabla} u\|_{L^2(\Omega)}} \leqslant C\left(\frac{\|\widehat{\nabla} u\|_{L^2(B_\rho(x_i))}}{\|\widehat{\nabla} u\|_{L^2(\Omega)}}\right)^{\delta}, \tag{4.10}$$

where $C > 0$ and δ, $0 < \delta < 1$, depend only on α_0, β_0, and M. By induction we have

$$\frac{\|\widehat{\nabla} u\|_{L^2(B_\rho(y))}}{\|\widehat{\nabla} u\|_{L^2(\Omega)}} \leqslant C^{1/(1-\delta)}\left(\frac{\|\widehat{\nabla} u\|_{L^2(B_\rho(x))}}{\|\widehat{\nabla} u\|_{L^2(\Omega)}}\right)^{\delta^L}. \tag{4.11}$$

Let us notice that $L \leqslant \frac{|\Omega|}{\omega_n\rho^n}$.

Let us cover $\Omega_{\frac{5\rho}{\theta}}$ with internally nonoverlapping closed cubes of side $l = 2\rho/\sqrt{n}\overline{\theta}$. Clearly, any such cube is contained in a ball of radius ρ and center in $\Omega_{\frac{4\rho}{\theta}}$ and the number of such cubes is controlled by $N = \frac{|\Omega| n^{n/2}\overline{\theta}^n}{2^n\rho^n}$. Therefore, from (4.11) we have

$$\frac{\|\widehat{\nabla} u\|_{L^2(\Omega_{\frac{5\rho}{\theta}})}}{\|\widehat{\nabla} u\|_{L^2(\Omega)}} \leqslant \frac{C}{\rho^{n/2}}\left(\frac{\|\widehat{\nabla} u\|_{L^2(B_\rho(x))}}{\|\widehat{\nabla} u\|_{L^2(\Omega)}}\right)^{\delta^L}, \tag{4.12}$$

where C depends only on $\alpha_0, \beta_0, M, |\Omega|$.

Now, let us estimate from below the left-hand side of (4.12) by means of a positive constant. Let us set

$$\frac{\|\widehat{\nabla} u\|^2_{L^2(\Omega_{\frac{5\rho}{\theta}})}}{\|\widehat{\nabla} u\|^2_{L^2(\Omega)}} = 1 - \frac{\int_{\Omega \setminus \Omega_{\frac{5\rho}{\theta}}} |\widehat{\nabla} u|^2}{\int_\Omega |\widehat{\nabla} u|^2}. \tag{4.13}$$

By a trace inequality (see, for instance, [LM72]) and by the Korn inequality (3.9), we have

$$\|\varphi\|_{H^{-1/2}(\partial\Omega)} \leqslant C \|\widehat{\nabla} u\|_{L^2(\Omega)}, \tag{4.14}$$

where C depends only on $\alpha_0, \beta_0, r_0, M_0, |\Omega|$. Hence, by (4.2) and (4.14), we have that there exists $\overline{\rho} > 0$, depending only on $\alpha_0, \beta_0, M, r_0, M_0, |\Omega|$, and $\|\varphi\|_{L^2(\partial\Omega)} / \|\varphi\|_{H^{-1/2}(\partial\Omega)}$, such that

$$\frac{\|\widehat{\nabla} u\|^2_{L^2(\Omega_{\frac{5\rho}{\theta}})}}{\|\widehat{\nabla} u\|^2_{L^2(\Omega)}} \geqslant \frac{1}{2} \tag{4.15}$$

for every ρ, $0 < \rho \leqslant \overline{\rho}$.

Finally, from (4.12) and (4.15) the thesis follows when $0 < \rho \leqslant \overline{\rho}$; for larger values of ρ, inequality (4.1) is trivial. □

PROPOSITION 4.3. *There exists θ^*, $0 < \theta^* \leqslant 1$, depending only on α_0, β_0, M, such that for every $\overline{r} > 0$ and every $x_0 \in \Omega_{\overline{r}}$ we have*

$$\int_{B_{2r}(x_0)} |u|^2 \leqslant K \int_{B_r(x_0)} |u|^2 \textit{ for every } r,\ 0 < r \leqslant \frac{\theta^* \overline{r}}{2}, \tag{4.16}$$

$$\int_{B_{2r}(x_0)} |\widehat{\nabla} u|^2 \leqslant K \int_{B_r(x_0)} |\widehat{\nabla} u|^2 \textit{ for every } r,\ 0 < r \leqslant \frac{\theta^* \overline{r}}{4}, \tag{4.17}$$

where $K > 0$ depends only on α_0, β_0, M, r_0, M_0, $|\Omega|$, $\overline{r}$, and $\|\varphi\|_{L^2(\partial\Omega)} / \|\varphi\|_{H^{-1/2}(\partial\Omega)}$, and θ^* is the quantity introduced in Corollary 3.10.

Proof. The proofs of (4.16) and (4.17) are similar. Let us prove (4.17), which takes a little bit more work. Given $x_0 \in \Omega_{\overline{r}}$ and r, $0 < r < \frac{\theta^* \overline{r}}{4}$, we may apply Corollary 3.10 with $R = \overline{r}$, obtaining (4.17) with K depending only on α_0, β_0, M, and

$$\tilde{N}_0(v; \theta^* \overline{r}) = \frac{(\theta^* \overline{r})^2 \int_{B_{\theta^* \overline{r}}(x_0)} |\nabla v|^2 + \overline{r}^2 |\nabla (\operatorname{div} v)|^2}{\int_{B_{\theta^* \overline{r}}(x_0)} |v|^2 + \overline{r}^2 |\operatorname{div} v|^2}, \tag{4.18}$$

the dependence on this last variable being monotonically increasing and where v is defined in $B_{\overline{r}}(x_0)$ by

$$v = u - c - W(x - x_0), \tag{4.19}$$

with

$$c = \frac{1}{|B_r(x_0)|} \int_{B_r(x_0)} u, \quad W = \frac{1}{2|B_r(x_0)|} \int_{B_r(x_0)} \nabla u - (\nabla u)^T. \tag{4.20}$$

We have that $\nabla v = \nabla u - W$, $\widehat{\nabla} v = \widehat{\nabla} u$, and $\operatorname{div} v = \operatorname{div} u$. Moreover, by interior regularity estimates (see [ADN64]), we have

$$\begin{aligned}(4.21)\quad |W| &\leqslant \frac{1}{|B_r(x_0)|}\int_{B_r(x_0)} |\nabla u| \leqslant \|\nabla u\|_{L^\infty(B_r(x_0))} \\ &\leqslant C\|u\|_{H^1(\Omega)} \leqslant C\|\varphi\|_{H^{-1/2}(\partial\Omega)},\end{aligned}$$

where C depends only on α_0, β_0, M, r_0, and $\overline{r}$. Hence

$$(4.22)\qquad \int_{B_{\theta^*\overline{r}}(x_0)} |\nabla v|^2 \leqslant 2\int_{B_{\theta^*\overline{r}}(x_0)} |\nabla u|^2 + |W|^2 \leqslant C\|\varphi\|^2_{H^{-1/2}(\partial\Omega)},$$

where C depends only on α_0, β_0, M, r_0, and $\overline{r}$. By the Caccioppoli-type inequality (3.6) we have

$$(4.23)\qquad \int_{B_{\theta^*\overline{r}}(x_0)} |v|^2 \geqslant C\int_{B_{\frac{\theta^*\overline{r}}{2}}(x_0)} |\nabla v|^2 \geqslant C\int_{B_{\frac{\theta^*\overline{r}}{2}}(x_0)} |\widehat{\nabla} u|^2,$$

where C depends only on α_0, β_0, M, $\overline{r}$. If $\frac{\theta^*}{2} \leqslant \frac{\overline{\theta}}{4}$, we may apply Proposition 4.1 with $\rho = \frac{\theta^*\overline{r}}{2}$, whereas if $\frac{\theta^*}{2} \geqslant \frac{\overline{\theta}}{4}$, we may apply Proposition 4.1 with $\rho = \frac{\overline{\theta}\overline{r}}{4}$. In both cases by trace theorems and the standard Korn inequality we obtain

$$(4.24)\qquad \int_{B_{\frac{\theta^*\overline{r}}{2}}(x_0)} |\widehat{\nabla} u|^2 \geqslant C\int_\Omega |\widehat{\nabla} u|^2 \geqslant C\|\varphi\|^2_{H^{-1/2}(\partial\Omega)},$$

where C depends only on α_0, β_0, M, r_0, M_0, $|\Omega|$, $\overline{r}$, and $\|\varphi\|_{L^2(\partial\Omega)}/\|\varphi\|_{H^{-1/2}(\partial\Omega)}$. By (4.22)–(4.24) and interior regularity estimates (see [ADN64]) we have

$$(4.25)\qquad \tilde{N}_0(v; \theta^*\overline{r}) \leqslant C,$$

where C depends only on α_0, β_0, M, r_0 M_0,$|\Omega|$, $\overline{r}$, and $\|\varphi\|_{L^2(\partial\Omega)}/\|\varphi\|_{H^{-1/2}(\partial\Omega)}$. Hence the thesis follows. □

Proposition 4.4 (A_p property). *For every $\overline{r} > 0$ there exist $B > 0$ and $p > 1$ such that for every $x_0 \in \Omega_{\overline{r}}$ we have*

$$(4.26)\quad \left(\frac{1}{|B_r(x_0)|}\int_{B_r(x_0)} |\widehat{\nabla} u|^2\right)\left(\frac{1}{|B_r(x_0)|}\int_{B_r(x_0)} |\widehat{\nabla} u|^{-2/(p-1)}\right)^{p-1} \leqslant B \qquad \text{for every } r,\ 0 < r \leqslant \frac{\theta^*\overline{r}}{4},$$

where θ^* is the quantity introduced in Corollary 3.10 and where B, p depend only on α_0, β_0, M, r_0, M_0,$|\Omega|$, $\overline{r}$, and $\|\varphi\|_{L^2(\partial\Omega)}/\|\varphi\|_{H^{-1/2}(\partial\Omega)}$.

Proof. In view of the results in [CF74] it is enough to prove a reverse Hölder's inequality for $|\widehat{\nabla} u|^2$. Let $v = u - c - W(x - x_0)$, with $c = \frac{1}{|B_{2r}(x_0)|}\int_{B_{2r}(x_0)} u$, $W = \frac{1}{2|B_{2r}(x_0)|}\int_{B_{2r}(x_0)} \nabla u - (\nabla u)^T$. By interior regularity estimates, the Korn inequality (3.9), and Proposition 4.3 we have

$$\begin{aligned}(4.27)\quad \|\widehat{\nabla} u\|_{L^\infty(B_r(x_0))} = \|\widehat{\nabla} v\|_{L^\infty(B_r(x_0))} &\leqslant \frac{C}{r^{(n+2)/2}}\|v\|_{H^1(B_{2r}(x_0))} \\ &\leqslant \frac{C}{r^{n/2}}\|\widehat{\nabla} u\|_{L^2(B_{2r}(x_0))} \leqslant \frac{C}{r^{n/2}}\|\widehat{\nabla} u\|_{L^2(B_r(x_0))},\end{aligned}$$

where C depends only on α_0, β_0, M, r_0, M_0,$|\Omega|$, $\overline{r}$, and $\|\varphi\|_{L^2(\partial\Omega)}/\|\varphi\|_{H^{-1/2}(\partial\Omega)}$. □

5. Proofs of Theorems 2.3 and 2.4. We base the proof of our main theorems on two auxiliary lemmas.

LEMMA 5.1. *Let the elasticity tensor fields* $\mathbb{C}(x)$ *and* $\widetilde{\mathbb{C}}(x)$ *satisfy* (2.5), (2.6) *in* Ω. *Suppose that weak solutions* u, $u_0 \in H^1(\Omega, \mathbb{R}^n)$ *to the traction problems* (2.25)–(2.26), (2.27)–(2.28) *exist. The following identities hold:*

$$\int_\Omega (\chi_{\Omega\setminus D}\mathbb{C} + \chi_D\widetilde{\mathbb{C}})\nabla(u-u_0)\cdot\nabla(u-u_0) - \int_D (\widetilde{\mathbb{C}}-\mathbb{C})\nabla u_0\cdot\nabla u_0 = \int_{\partial\Omega}(g-g_0)\cdot\varphi, \tag{5.1}$$

$$\int_\Omega \mathbb{C}\nabla(u-u_0)\cdot\nabla(u-u_0) + \int_D (\widetilde{\mathbb{C}}-\mathbb{C})\nabla u\cdot\nabla u = \int_{\partial\Omega}(g_0-g)\cdot\varphi, \tag{5.2}$$

$$\int_D (\widetilde{\mathbb{C}}-\mathbb{C})\nabla u\cdot\nabla u_0 = \int_{\partial\Omega}(g_0-g)\cdot\varphi, \tag{5.3}$$

where g,$g_0 \in H^{1/2}(\partial\Omega, \mathbb{R}^n)$ *are the traces of* u,u_0, *respectively, on* $\partial\Omega$.

Proof of Lemma 5.1. Let us denote $\mathbb{H} = (\widetilde{\mathbb{C}} - \mathbb{C})$ in Ω. From the weak formulation of the problem (2.25)–(2.26) with $D = D_1$ and $D = D_2$ we get the identity

$$\int_\Omega (\mathbb{C} + \chi_{D_1}\mathbb{H})\nabla u_1\cdot\nabla w = \int_\Omega (\mathbb{C} + \chi_{D_2}\mathbb{H})\nabla u_2\cdot\nabla w \quad \text{for every} \quad w\in H^1(\Omega,\mathbb{R}^n), \tag{5.4}$$

where u_i is the solution to (2.25)–(2.26) with $D = D_i$, $i = 1, 2$, respectively. Subtracting the quantity $\int_\Omega (\mathbb{C} + \chi_{D_1}\mathbb{H})\nabla u_2\cdot\nabla w$ from both sides of (5.4) we have

$$\int_\Omega (\mathbb{C} + \chi_{D_1}\mathbb{H})\nabla(u_1-u_2)\cdot\nabla w = \int_\Omega (\chi_{D_2} - \chi_{D_1})\mathbb{H}\nabla u_2\cdot\nabla w \quad \text{for every} \quad w\in H^1(\Omega,\mathbb{R}^n). \tag{5.5}$$

Choosing $w = u_1$ into (5.5) we get

$$\int_\Omega (\mathbb{C} + \chi_{D_1}\mathbb{H})\nabla(u_1-u_2)\cdot\nabla u_1 = \int_\Omega (\chi_{D_2} - \chi_{D_1})\mathbb{H}\nabla u_2\cdot\nabla u_1. \tag{5.6}$$

By using the weak formulation of the traction problems for u_1 and u_2, the left-hand side of (5.6) can be rewritten as follows:

$$\int_\Omega (\mathbb{C} + \chi_{D_1}\mathbb{H})\nabla(u_1-u_2)\cdot\nabla u_1 = \int_{\partial\Omega}(g_1-g_2)\cdot\varphi, \tag{5.7}$$

where $g_i = u_i\,|_{\partial\Omega}$, $i = 1, 2$, and therefore (5.6) becomes

$$\int_{\partial\Omega}(g_1-g_2)\cdot\varphi = \int_\Omega (\chi_{D_2} - \chi_{D_1})\mathbb{H}\nabla u_2\cdot\nabla u_1. \tag{5.8}$$

Choosing $w = u_1 - u_2$ into (5.5) and using (5.8) we get

$$\int_\Omega (\mathbb{C} + \chi_{D_1}\mathbb{H})\nabla(u_1 - u_2) \cdot \nabla(u_1 - u_2) = \int_\Omega (\chi_{D_1} - \chi_{D_2})\mathbb{H}\nabla u_2 \cdot \nabla u_2 + \int_{\partial\Omega} (g_1 - g_2) \cdot \varphi, \tag{5.9}$$

and finally we obtain the fundamental identity

$$\int_\Omega (\mathbb{C} + \chi_{D_1}\mathbb{H})\nabla(u_1 - u_2) \cdot \nabla(u_1 - u_2) + \int_{D_2 \setminus D_1} \mathbb{H}\nabla u_2 \cdot \nabla u_2 = \int_{\partial\Omega} (g_1 - g_2) \cdot \varphi + \int_{D_1 \setminus D_2} \mathbb{H}\nabla u_2 \cdot \nabla u_2, \tag{5.10}$$

which is the analogue to the identity found in Kang, Seo, and Sheen [KSS97] for the inverse conductivity problem.

By choosing $D_1 = D$ and $D_2 = \emptyset$ we obtain the first identity (5.1) of the lemma. The second identity (5.2) follows from (5.10) for $D_1 = \emptyset$ and $D_2 = D$.

To get the third identity (5.3), we choose $w = u_0$ and $w = u$ in the weak formulation of the traction problem (2.25)–(2.26) for $D = D_1$ and $D = \emptyset$, respectively:

$$\int_\Omega (\mathbb{C} + \chi_D\mathbb{H})\nabla u \cdot \nabla u_0 = \int_{\partial\Omega} g_0 \cdot \varphi, \tag{5.11}$$

$$\int_\Omega \mathbb{C}\nabla u_0 \cdot \nabla u = \int_{\partial\Omega} g \cdot \varphi. \tag{5.12}$$

Subtracting (5.12) from (5.11) we obtain identity (5.3). □

Lemma 5.2. *Let $\mathbb{C}(x)$ and $\widetilde{\mathbb{C}}(x)$ satisfy* (2.5), (2.6) *in Ω. Let ξ_0, ξ_1, $0 < \xi_0 < \xi_1$, be such that*

$$\xi_0|A|^2 \le \mathbb{C}(x)A \cdot A \le \xi_1|A|^2 \qquad for \quad a.e. \quad x \in \Omega, \tag{5.13}$$

for any symmetric $n \times n$ matrix A, and let the jump $(\widetilde{\mathbb{C}}(x) - \mathbb{C}(x))$ satisfy either (2.30) *or* (2.31)*. Let u, $u_0 \in H^1(\Omega, \mathbb{R}^n)$ be the weak solutions to the traction problems* (2.25)–(2.26), (2.27)–(2.28)*, respectively.*

If (2.30) *holds, then we have*

$$\frac{\eta\xi_0}{\delta} \int_D |\widehat{\nabla} u_0|^2 \le \int_{\partial\Omega} (g_0 - g) \cdot \varphi \le (\delta - 1)\xi_1 \int_D |\widehat{\nabla} u_0|^2. \tag{5.14}$$

If instead (2.31) *holds, then we have*

$$\eta\xi_0 \int_D |\widehat{\nabla} u_0|^2 \le \int_{\partial\Omega} (g - g_0) \cdot \varphi \le \frac{1 - \delta}{\delta}\xi_1 \int_D |\widehat{\nabla} u_0|^2. \tag{5.15}$$

Proof of Lemma 5.2. Suppose that (2.30) holds. Then, from identity (5.1) we have

$$\int_{\partial\Omega} (g_0 - g) \cdot \varphi \le \int_D \mathbb{H}\nabla u_0 \cdot \nabla u_0, \tag{5.16}$$

where $\mathbb{H} = (\widetilde{\mathbb{C}} - \mathbb{C})$ in Ω. The inequality below follows by the symmetry properties (2.7), (2.8), (2.9) and the positivity condition (2.30):

$$\int_D \mathbb{H}\nabla u_0 \cdot \nabla u_0 \leq (1+\epsilon)\int_D \mathbb{H}\nabla(u-u_0)\cdot\nabla(u-u_0) + \left(1+\frac{1}{\epsilon}\right)\int_D \mathbb{H}\nabla u\cdot\nabla u \qquad \text{for every} \quad \epsilon > 0. \tag{5.17}$$

Then, from (2.30) we have

$$\int_D \mathbb{H}\nabla u_0 \cdot \nabla u_0 \leq (1+\epsilon)(\delta-1)\left[\int_D \mathbb{C}\nabla(u-u_0)\cdot\nabla(u-u_0) + \frac{1}{\epsilon(\delta-1)}\int_D \mathbb{H}\nabla u\cdot\nabla u\right]. \tag{5.18}$$

Choosing $\epsilon = \frac{1}{\delta-1}$ in (5.18) and by employing identity (5.2) we get

$$\int_D \mathbb{H}\nabla u_0 \cdot \nabla u_0 \leq \delta \int_{\partial\Omega}(g_0 - g)\cdot\varphi. \tag{5.19}$$

The double inequality (5.14) follows from (5.16), (5.19) and (5.13), (2.30).

In the case where (2.31) holds, from (5.1) we have

$$\int_{\partial\Omega}(g-g_0)\cdot\varphi \geq \int_D -\mathbb{H}\nabla u_0\cdot\nabla u_0. \tag{5.20}$$

From (5.3) we obtain $\int_{\partial\Omega}(g-g_0)\cdot\varphi = \int_D -\mathbb{H}\nabla u\cdot\nabla u_0$, and then, reasoning as in (5.17), we find

$$\int_{\partial\Omega}(g-g_0)\cdot\varphi \leq \frac{\epsilon}{2}\int_D -\mathbb{H}\nabla u\cdot\nabla u + \frac{1}{2\epsilon}\int_D -\mathbb{H}\nabla u_0\cdot\nabla u_0 \qquad \text{for every} \quad \epsilon > 0. \tag{5.21}$$

By using (5.2), (2.31), and (5.1) we have

$$\begin{aligned}\int_D -\mathbb{H}\nabla u\cdot\nabla u &= \int_{\partial\Omega}(g-g_0)\cdot\varphi + \int_\Omega \mathbb{C}\nabla(u-u_0)\cdot\nabla(u-u_0)\\ &\leq \int_{\partial\Omega}(g-g_0)\cdot\varphi + \frac{1}{\delta}\int_\Omega(\mathbb{C}+\mathbb{H})\nabla(u-u_0)\cdot\nabla(u-u_0)\\ &= \int_{\partial\Omega}(g-g_0)\cdot\varphi + \frac{1}{\delta}\left[\int_{\partial\Omega}(g-g_0)\cdot\varphi + \int_\Omega \mathbb{H}\nabla u_0\cdot\nabla u_0\right]\\ &= \frac{\delta+1}{\delta}\int_{\partial\Omega}(g-g_0)\cdot\varphi + \frac{1}{\delta}\int_{\mathbb{D}} \mathbb{H}\nabla u_0\cdot\nabla u_0.\end{aligned} \tag{5.22}$$

Inserting inequality (5.22) into (5.21), we obtain

$$\int_{\partial\Omega}(g-g_0)\cdot\varphi \leq \alpha(\epsilon)\int_D -\mathbb{H}\nabla u_0\cdot\nabla u_0, \tag{5.23}$$

where $\alpha(\epsilon) = \frac{\delta-\epsilon^2}{\epsilon(2\delta-\epsilon\delta-\epsilon)}$. The minimum of $\alpha(\epsilon)$ occurs when $\epsilon = \delta$ and in this case we have

$$\int_{\partial\Omega} (g-g_0)\cdot\varphi \leq \frac{1}{\delta}\int_D -\mathbb{H}\nabla u_0\cdot\nabla u_0. \tag{5.24}$$

The double inequality (5.15) follows from (5.20), (5.24) and (5.13), (2.31). □

Proof of Theorem 2.3. By (2.21), the inequality (5.13) holds, with $\xi_0 = \min\{2\alpha_0, \gamma_0\}$, $\xi_1 = (n+2)M$, so that Lemma 5.2 may be applied.

By standard regularity estimates for elliptic systems (see Agmon, Douglis, and Nirenberg [ADN64]), by the Korn inequality, by (5.13), and by the weak formulation of the Neumann problem (2.27)-(2.28), we have

$$\|\widehat{\nabla}u_0\|_{L^\infty(D)} \leq C\|u_0\|_{H^1(\Omega)} \leq C\|\widehat{\nabla}u_0\|_{L^2(\Omega)} \leq C\left(\int_{\partial\Omega} g_0\cdot\varphi\right)^{\frac{1}{2}}, \tag{5.25}$$

where the constant C depends only on α_0, γ_0, M, d_0, and $|\Omega|$.

The lower bound for $|D|$ in (2.34), (2.35) follows from the right-hand side of (5.14), (5.15) and from (5.25).

Let us prove the upper bound for $|D|$ in (2.34), (2.35). Let $\epsilon = \min\{\frac{\bar{\theta}d_0}{2}, \frac{h_1}{\sqrt{n}}\}$, where $\bar{\theta}$ is as in Proposition 4.1. Let us cover D_{h_1} with internally nonoverlapping closed cubes Q_l of side ϵ, for $l = 1,\dots,L$. By the choice of ϵ the cubes Q_l are contained in D. Hence

$$\int_D \left|\widehat{\nabla}u_0\right|^2 \geq \int_{\bigcup_{l=1}^L Q_l} \left|\widehat{\nabla}u_0\right|^2 \geq \frac{|D_{h_1}|}{\epsilon^n}\int_{Q_{\bar{l}}} \left|\widehat{\nabla}u_0\right|^2, \tag{5.26}$$

where $\bar{l}$ is such that $\int_{Q_{\bar{l}}} \left|\widehat{\nabla}u_0\right|^2 = \min_l \int_{Q_l} \left|\widehat{\nabla}u_0\right|^2$. Let $\bar{x}$ be the center of $Q_{\bar{l}}$. From (5.26), estimate (4.1) with $x = \bar{x}$ and $\rho = \epsilon/2$ from (5.13), and from the weak formulation of (2.27)-(2.28) we have

$$\int_D \left|\widehat{\nabla}u_0\right|^2 \geq K|D|\int_{\partial\Omega} g_0\cdot\varphi, \tag{5.27}$$

where K depends only on α_0, β_0, d_0, $|\Omega|$, r_0, M_0, M, h_1, $\|\varphi\|_{L^2(\partial\Omega)}/\|\varphi\|_{H^{-1/2}(\partial\Omega)}$. The upper bound for D in (2.34), (2.35) follows from the left-hand side of (5.14), (5.15) and from (5.27). □

Proof of Theorem 2.4. Let $\bar{r} = \frac{d_0}{2}$ and $\epsilon = \min\{\frac{\theta^* d_0}{4\sqrt{n}}, \frac{\bar{\theta}d_0}{4}\}$, where θ^* is as in Proposition 4.4. Let us cover D with internally nonoverlapping closed cubes Q_j, $j = 1,\dots,J$, with side ϵ. By Hölder's inequality we have

$$|D| \leq \left(\int_{\bigcup_{j=1}^J Q_j} \left|\widehat{\nabla}u_0\right|^{-\frac{2}{p-1}}\right)^{\frac{p-1}{p}} \left(\int_D \left|\widehat{\nabla}u_0\right|^2\right)^{\frac{1}{p}}, \tag{5.28}$$

where p is chosen as in Proposition 4.4. By applying Proposition 4.4 to the balls B_j circumscribing each Q_j, $j = 1,\dots,J$, we have

$$\left(\int_{\bigcup_{j=1}^J Q_j} |\widehat{\nabla} u_0|^{-\frac{2}{p-1}}\right)^{\frac{p-1}{p}} = \left(\epsilon^n \sum_{j=1}^J \frac{1}{|Q_j|}\int_{Q_j} |\widehat{\nabla} u_0|^{-\frac{2}{p-1}}\right)^{\frac{p-1}{p}}$$
$$\le \left(\epsilon^n \sum_{j=1}^J \left(\frac{B[C(n)]^p}{\frac{1}{|Q_j|}\int_{Q_j}|\widehat{\nabla} u_0|^2}\right)^{\frac{1}{p-1}}\right)^{\frac{p-1}{p}} \le \frac{(J\epsilon^n)^{\frac{p-1}{p}} B^{\frac{1}{p}} C(n)}{\min_j \left(\frac{1}{|Q_j|}\int_{Q_j}|\widehat{\nabla} u_0|^2\right)^{\frac{1}{p}}}, \tag{5.29}$$

where $C(n) = \omega_n \left(\frac{\sqrt{n}}{2}\right)^n$ and B is as in Proposition 4.4. Now $J\epsilon^n = \sum_{j=1}^J |Q_j| \le |\Omega|$ and hence, from (5.28), we have

$$|D| \le |\Omega|^{\frac{p-1}{p}} B^{\frac{1}{p}} C(n) \left(\frac{\epsilon^n \int_D |\widehat{\nabla} u_0|^2}{\min_j \int_{Q_j} |\widehat{\nabla} u_0|^2}\right)^{\frac{1}{p}}. \tag{5.30}$$

By Proposition 4.1, (5.30), (5.13), and the weak formulation of (2.27)-(2.28), we have

$$\int_D |\widehat{\nabla} u_0|^2 \ge \left(K \int_{\partial\Omega} g_0 \cdot \varphi\right) |D|^p, \tag{5.31}$$

where K depends only on α_0, β_0, d_0, $|\Omega|$, r_0, M_0, M, h_1, $\|\varphi\|_{L^2(\partial\Omega)}/\|\varphi\|_{H^{-1/2}(\partial\Omega)}$. The right-hand side of (2.36), (2.37) follow from the left-hand side of (5.14), (5.15), and (5.31). □

Acknowledgment. The authors wish to thank Mikhail Belishev, who kindly pointed out the reference [L86].

REFERENCES

[Ad75] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[ADN64] S. Agmon, A. Douglis, and N. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions,* II, Comm. Pure Appl. Math., 15 (1964), pp. 35–92.

[AI96] G. Alessandrini and V. Isakov, *Analyticity and uniqueness for the inverse conductivity problem*, Rend. Istit. Mat. Univ. Trieste, 28 (1996), pp. 351–370.

[AITY98] D. D. Ang, M. Ikehata, D. D. Trong, and M. Yamamoto, *Unique continuation for a stationary isotropic Lamé system with variable coefficients*, Comm. Partial Differential Equations, 23 (1998), pp. 371–385.

[AKS62] N. Aronszajn, A. Krzywicki, and J. Szarski, *A unique continuation theorem for exterior differential forms on Riemannian manifolds*, Ark. Mat., 4 (1962), pp. 417–453.

[Al99] G. Alessandrini, *Generic uniqueness and size estimates in the inverse conductivity problem with one measurement*, Matematiche (Catania), 54 (1999), pp. 5–14.

[AM01] G. Alessandrini and A. Morassi, *Strong unique continuation for the Lamé system of elasticity*, Comm. Partial Differential Equations, 26 (2001), pp. 1787–1810.

[AR98] G. Alessandrini and E. Rosset, *The inverse conductivity problem with one measurement: Bounds on the size of the unknown object*, SIAM J. Appl. Math., 58 (1998), pp. 1060–1071.

[ARS00] G. Alessandrini, E. Rosset, and J. K. Seo, *Optimal size estimates for the inverse conductivity problem with one measurement*, Proc. Amer. Math. Soc., 128 (2000), pp. 53–64.

[CF74] R. R. Coifman and C. L. Fefferman, *Weighted norm inequalities for maximal functions and singular integrals*, Studia Math., 51 (1974), pp. 241–250.

[DR93] B. Dehman and L. Robbiano, *La propriété du prolungement unique pour un système elliptique: le système de Lamé*, J. Math. Pures Appl., 72 (1993), pp. 475–492.

[EINT00] M. Eller, V. Isakov, G. Nakamura, and D. Tataru, *Uniqueness and stability in the Cauchy problem for Maxwell and elasticity systems,* in Nonlinear Partial Differential Equations, *College de France Seminar,* Vol. 14, Chapman and Hall/CRC Press, 2000.

[Fi72] G. Fichera, *Existence theorems in elasticity*, in Handbuch der Physik, Vol. VI, Springer-Verlag, Berlin, Heidelberg, New York, 1972, pp. 347–389.

[Fr87] A. Friedman, *Detection of mines by electric measurements*, SIAM J. Appl. Math., 47 (1987), pp. 201–212.

[FrG87] A. Friedman and B. Gustafsson, *Identification of the conductivity coefficient in an elliptic equation*, SIAM J. Math. Anal., 18 (1987), pp. 777–787.

[FrI89] A. Friedman and V. Isakov, *On the uniqueness in the inverse conductivity problem with one measurement*, Indiana Univ. Math. J., 38 (1989), pp. 563–579.

[GCRDF85] J. Garcia-Cuerva and J. L. Rubio De Francia, *Weighted Norm Inequalities and Related Topics*, North–Holland, Amsterdam, 1985.

[GL86] N. Garofalo and F. H. Lin, *Monotonicity properties of variational integrals, A_p weights and unique continuation*, Indiana Univ. Math. J., 35 (1986), pp. 245–268.

[GL87] N. Garofalo and F. H. Lin, *Unique continuation for elliptic operators: A geometric-variational approach,* Comm. Pure Appl. Math., 40 (1987), pp. 347–366.

[Gur72] M. E. Gurtin, *The linear theory of elasticity*, in Handbuch der Physik, Vol. VI, Springer-Verlag, Berlin, Heidelberg, New York, 1972, pp. 1–295.

[I98] M. Ikehata, *Size estimation of inclusion*, J. Inverse Ill-Posed Probl., 6 (1998), pp. 127–140.

[KSS97] H. Kang, J. K. Seo, and D. Sheen, *The inverse conductivity problem with one measurement: Stability and estimation of size*, SIAM J. Math. Anal., 28 (1997), pp. 1389–1405.

[KT01] H. Koch and D. Tataru, *Carleman estimates and unique continuation for second-order elliptic equations with nonsmooth coefficients*, Comm. Pure Appl. Math., 54 (2001), pp. 339–360.

[L86] R. Leis, *Initial-Boundary Value Problems in Mathematical Physics*, B. G. Teubner, Stuttgart, Germany, John Wiley and Sons, Chichester, UK, 1986.

[LM72] J. L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications,* Vol. I, Springer-Verlag, Berlin, 1972.

[MS58] E. Magenes and G. Stampacchia, *I problemi al contorno per le equazioni differenziali di tipo ellittico*, Ann. Scuola Norm. Sup. Pisa (3), 12 (1958), pp. 247–358.

[T99] A. Tiero, *On Korn's inequality in the second case*, J. Elasticity, 54 (1999), pp. 187–191.

[V88] T. Valent, *Boundary Value Problems of Finite Elasticity*, Springer-Verlag, New York, 1988.

[W69] N. Weck, *Außenraumaufgaben in der Theorie stationärer Schwingungen inhomogener elasticher Körper*, Math. Z., 111 (1969), pp. 387–398.

[W01] N. Weck, *Unique continuation for systems with Lamé principal part*, Math. Methods Appl. Sci., 24 (2001), pp. 595–605.

SIAM J. MATH. ANAL.
Vol. 33, No. 6, pp. 1269–1296

THE GLOBAL CAUCHY PROBLEM IN BOURGAIN'S-TYPE SPACES FOR A DISPERSIVE DISSIPATIVE SEMILINEAR EQUATION*

LUC MOLINET† AND FRANCIS RIBAUD‡

Abstract. We prove local and global well-posedness results for the Kadomtsev–Petviashvili–Burgers equations in Bourgain's-type spaces. This approach is new for the study of semilinear evolution equations with a linear part which contains both dispersive and dissipative terms.

Key words. Kadomtsev–Petviashvili–Burgers equations, Bourgain spaces, global existence

AMS subject classifications. 35Q53, 35M10, 35G25

PII. S0036141000374634

1. Introduction.

We study the initial value problem for the Kadomtsev–Petviashvili–Burgers (KPB) equations in $\mathbb{R}^2$:

$$\begin{cases} (u_t + u_{xxx} + uu_x - u_{xx})_x + \varepsilon u_{yy} = 0, \\ u(0) = \varphi, \end{cases} \tag{1.1}$$

where $\varepsilon = \pm 1$.

These equations are dissipative versions of the Kadomtsev–Petviashvili (KP) equations

$$(u_t + u_{xxx} + uu_x)_x + \varepsilon u_{yy} = 0,$$

which are "universal" models for nearly one directional weakly nonlinear dispersive waves with weak transverse effects. The KP equations being themselves natural two dimensional extensions of the famous KdV equation

$$u_t + u_{xxx} + uu_x = 0.$$

In some physical contexts (such as magnetosonic waves damped by electron-ion collisions), when dissipative effects cannot be neglected, the KdV–Burgers equation

$$u_t + u_{xxx} + uu_x - u_{xx} = 0$$

has been derived (cf. [14]). The KPB equations are thus natural candidates to model the propagation of two dimensional damped waves in some physical situations. Note that due to the hypothesis of nearly one directional propagation, the dissipative term acts only in the main direction of propagation in (1.1).

In the last decade, Bourgain developed a new method to study the Cauchy problem for nonlinear dispersive equations. This method was successfully applied to the

*Received by the editors June 28, 2000; accepted for publication (in revised form) November 6, 2001; published electronically April 26, 2002.
http://www.siam.org/journals/sima/33-6/37463.html

†L.A.G.A., Institut Galilée, Université Paris 13, 93430 Villetaneuse, France (luc.molinet@math.u-psud.fr).

‡Equipe d'Analyse et de Mathématiques Appliquées, Université de Marne-la-Vallée, Cité Descartes, 5 Bd. Descartes, Champs sur Marne, 77454 Marne-la-Vallée Cedex 2, France (ribaud@math.univ-mlv.fr).

Schrödinger equation, the KdV equation, and the KP II equation ($\varepsilon = +1$). Concerning the KdV and KP II equations, it permitted, for instance, to show the global well-posedness in H^s, $s \geq 0$; see, respectively, [4] and [3]. Also local existence results in Sobolev spaces of negative order have been obtained; see [9] and [16]. One of the particularities of this method is its use of special Fourier transform restriction spaces strongly related to the symbol of the linear equation, the existence and uniqueness of the solution then being obtained by a standard fixed point argument on the corresponding integral equation.

The main contribution of this paper is to show how the Bourgain spaces can also be successfully employed to study the Cauchy problem for semilinear equations with a linear part which contains both dissipative and dispersive terms. As far as the authors know, Bourgain spaces were never used in such a context.

Let us recall that in the Bourgain's method for the KP II equation, the loss of derivative in the nonlinear term $\partial_x(u^2)$ is compensated by a smoothing effect resulting from a simple algebraic inequality (see (5.2)) involving the symbol $\tau - \xi^3 - \eta^2/\xi$ of the linear KP II equation.

In the case of the KP I equation ($\varepsilon = -1$), this inequality does not hold, which is why there is no available result on the local well-posedness of the KP I equation by using Bourgain spaces. However, by the parabolic regularization method, local well-posedness in $H^s(\mathbb{R}^2)$, $s > 2$, has been obtained (cf. [5]). But this approach seems far from satisfactory since it does not use at all the dispersive nature of the equation.

In this paper the main idea is to work in the Bourgain spaces associated with the "usual" KP equations, i.e., related only to the dispersive part of the linear symbol of (1.1) (see Remark 3 below). After having extended the linear semigroup of the KPB equations to a linear operator $W(\cdot)$ defined on the whole real axis, we first derive some linear estimates in the Bourgain spaces for the "free evolution term" $W(t)\varphi$ and for the "forcing term" L defined by (2.10). In contrast to the purely dispersive case, we show that L is a smoothing operator not only in time but also in space. Then by using Strichartz's type estimates for the KP equations injected into the framework of Bourgain spaces, we show a bilinear estimate which will yield the local well-posedness of the KPB equations ($\varepsilon = +1$ or -1) in the anisotropic Sobolev spaces $H^{s_1,s_2}(\mathbb{R}^2)$, provided $s_1 > 0$ and $s_2 \geq 0$. Moreover, combining these Strichartz's estimates and the algebraic inequality (5.2), we prove the local well-posedness in $H^s(\mathbb{R}^2)$, $s \geq 0$, for the KPB II equation ($\varepsilon = +1$). The global well-posedness results will then follow by means of a priori estimates.

Let us note that local well-posedness for the KPB equations in the space $\{\varphi \in H^1(\mathbb{R}^2),\ \mathcal{F}^{-1}_{x,y}(\frac{\eta}{\xi}\hat{\varphi}) \in L^2(\mathbb{R}^2)\}$ was obtained in [12]. Also, global existence was obtained under a smallness restriction on the initial data when $\varepsilon = +1$. Finally, a local smoothing effect in the transverse direction was proved as well as the existence of global weak solutions in $L^2(\mathbb{R}^2)$ when $\varepsilon = -1$.

1.1. Notations. For $f \in \mathcal{S}'$ we denote by $\hat{f}$ or $\mathcal{F}(f)$ the Fourier transform of f, i.e.,

$$\hat{f}(\theta) = \int_{\mathbb{R}^n} e^{-i\langle \alpha, \theta\rangle} f(\alpha)\, d\alpha,$$

and we define the linear operator Δ_x^b by

$$\Delta_x^b(f)(x,y) = \mathcal{F}^{-1}\Big((1+|\xi|^2)^b \hat{f}(\xi,\eta)\Big)(x,y).$$

For a Banach space X, we denote by $\|\cdot\|_X$ the norm in X. We will use the Sobolev spaces $H^s(\mathbb{R}^2)$ and the homogeneous Sobolev spaces $\dot{H}^s(\mathbb{R}^2)$ equipped with the norms

$$\|u\|_{H^s}^2=\int_{\mathbb{R}^2}\langle\zeta\rangle^{2s}|\hat{u}(\zeta)|^2\,d\zeta,\qquad \|u\|_{\dot{H}^s}^2=\int_{\mathbb{R}^2}|\zeta|^{2s}|\hat{u}(\zeta)|^2\,d\zeta,$$

where $\zeta=(\xi,\eta)$, $|\zeta|=(|\xi|^2+|\eta|^2)^{1/2}$, and $\langle\zeta\rangle=(1+|\zeta|^2)^{1/2}$.

We also consider the anisotropic Sobolev spaces $H^{s_1,s_2}(\mathbb{R}^2)$ endowed with the norm

$$\|u\|_{H^{s_1,s_2}}^2=\int_{\mathbb{R}^2}\langle\xi\rangle^{2s_1}\langle\eta\rangle^{2s_2}\,|\hat{u}(\zeta)|^2\,d\zeta.$$

Next we consider the corresponding space-time Sobolev spaces $H^{b,s}$ and H^{b,s_1,s_2}, respectively, equipped with the norms

$$\|u\|_{H^{b,s}}^2=\int_{\mathbb{R}^3}\langle\tau\rangle^{2b}\langle\zeta\rangle^{2s}\,|\hat{u}(\tau,\zeta)|^2\,d\zeta\,d\tau,$$

$$\|u\|_{H^{b,s_1,s_2}}^2=\int_{\mathbb{R}^3}\langle\tau\rangle^{2b}\langle\xi\rangle^{2s_1}\langle\eta\rangle^{2s_2}\,|\hat{u}(\tau,\zeta)|^2\,d\zeta\,d\tau.$$

We will also use the space-time Lebesgue spaces $L^{p,q}_{t,z}$ $(z=(x,y))$ endowed with the norm

$$\|u\|_{L^{q,r}_{t,z}}=\big\|\|u\|_{L^r_z}\big\|_{L^q_t},$$

and we will use the notation $L^2_{t,z}$ for $L^{2,2}_{t,z}$.

Let $U(\cdot)$ be the unitary group which defined the free evolution of the KP equation, i.e.,

$$U(t)=\exp(itP(D_x,D_y)), \tag{1.2}$$

where $P(D_x,D_y)$ is the Fourier multiplier with symbol

$$P(\xi,\eta)=\xi^3-\varepsilon\,\frac{\eta^2}{\xi},\quad \varepsilon=\pm1.$$

We denote by $X^{b,s}$ and X^{b,s_1,s_2} the Bourgain's-type spaces associated with the spaces $H^{b,s}$ and H^{b,s_1,s_2} for the KP equations. They are, respectively, endowed with the norms

$$\|u\|_{X^{b,s}}=\|U(-t)u\|_{H^{b,s}} \tag{1.3}$$

and

$$\|u\|_{X^{b,s_1,s_2}}=\|U(-t)u\|_{H^{b,s_1,s_2}}. \tag{1.4}$$

Note that, since $\mathcal{F}(U(-t)u)(\tau,\zeta)=\mathcal{F}(u)(\tau+P(\zeta),\zeta)$, one can, respectively, re-express the norm of $X^{b,s}$ and X^{b,s_1,s_2} as

$$\|u\|_{X^{b,s}}=\|\langle\tau-P(\zeta)\rangle^b\,\langle\zeta\rangle^s\,\hat{u}(\tau,\zeta)\|_{L^2(\mathbb{R}^3)}$$

and

$$\|u\|_{X^{b,s_1,s_2}} = \|\langle \tau - P(\zeta)\rangle^b \, \langle \xi \rangle^{s_1} \langle \eta \rangle^{s_2} \, \hat{u}(\tau,\zeta)\|_{L^2(\mathbb{R}^3)}.$$

For $T \geq 0$, we consider the localized Bourgain spaces $X_T^{b,s}$ endowed with the norm

$$\|u\|_{X_T^{b,s}} = \inf_{w \in X^{b,s}} \{\|w\|_{X^{b,s}}, \, w(t) = u(t) \text{ on } [0,T] \,\},$$

the space X_T^{b,s_1,s_2} being defined in the same way.

Finally we denote by $W(\cdot)$ the semigroup associated with the free evolution of the KPB equations, i.e.,

$$\forall t \geq 0, \; \mathcal{F}_z(W(t)\varphi)(\zeta) = \exp[-|\xi|^2 t + iP(\zeta)\, t]\hat{\varphi}(\zeta), \qquad \varphi \in \mathcal{S}',$$

and we extend $W(\cdot)$ to a linear operator defined on the whole real axis by setting

$$\forall t \in \mathbb{R}, \; \mathcal{F}_z(W(t)\varphi)(\zeta) = \exp[-|\xi|^2 |t| + iP(\zeta) t]\hat{\varphi}(\zeta), \qquad \varphi \in \mathcal{S}'.$$

1.2. Main results. To prove local well-posedness results, we shall apply a fixed point argument in $X_T^{b,s}$ or X_T^{b,s_1,s_2} to the following cut-off version of the integral equation associated with (1.1):

$$u(t) = \psi(t) \left[W(t)\varphi - \frac{\mathbb{1}_{\mathbb{R}_+}(t)}{2} \int_0^t W(t-t')\partial_x(\psi_T^2(t')u^2(t'))\, dt' \right], \tag{1.5}$$

where $t \in \mathbb{R}$ and, in the rest of this paper, ψ is a time cut-off function satisfying

$$\psi \in C_0^\infty(\mathbb{R}), \quad \text{supp}\, \psi \subset [-1,1], \quad \psi = 1 \text{ on } \left[-\frac{1}{2}, \frac{1}{2}\right],$$

and $\psi_T(\cdot) = \psi(\cdot/T)$.

THEOREM 1.1 (KPB I). *Let $\varepsilon = -1$ and $(s_1,s_2) \in \mathbb{R}_+^* \times \mathbb{R}_+$. Then for any $\varphi \in H^{s_1,s_2}$, there exist a positive $T = T(\|\varphi\|_{H^{0+,0}})$ and a unique solution u to* (1.1) *in*

$$Y_T = C([0,T], \, H^{s_1,s_2}) \cap X_T^{1/2,s_1,s_2}. \tag{1.6}$$

Also, the map $\varphi \mapsto u$ is continuous from H^{s_1,s_2} to Y_T.

Moreover, if $s_1 \geq 1$ and $\mathcal{F}_z^{-1}(\frac{\eta}{\xi}\hat{\varphi}) \in L^2(\mathbb{R}^2)$, then T can be chosen arbitrarily large and

$$\forall t \geq 0, \; \|u(t)\|_{H^{1,0}} \leq C(\|\varphi\|_{L^2}, \|\partial_x \varphi\|_{L^2}, \|\partial_x^{-1}\varphi_y\|_{L^2}).$$

THEOREM 1.2 (KPB II). *Let $\varepsilon = 1$ and let $\varphi \in H^s(\mathbb{R}^2)$, $s \geq 0$. Then for any $T > 0$, there exists a unique solution u to* (1.1) *in*

$$Y_T = C([0,T], \, H^s) \cap X_T^{1/2,s}. \tag{1.7}$$

Moreover, the map $\varphi \mapsto u$ is continuous from $H^s(\mathbb{R}^2)$ to Y_T and the following inequality holds:

$$\forall t \geq 0, \; \|u(t)\|_{L^2} \leq \|\varphi\|_{L^2}.$$

Remark 1. It is easy to see that a solution of the integral equation (1.5) solves (1.1) in the distribution sense. Indeed, let $u \in C([0,T], H^s)$ such that, on $[0,T]$,

$$u(t) = W(t)\varphi - \frac{1}{2}\int_0^t W(t-t')\partial_x(u^2(t'))\,dt' . \tag{1.8}$$

Differentiating (1.8) with respect to x and then to t, we readily obtain

$$\partial_t\partial_x u + \frac{1}{2}\partial_x^2(u^2) + \partial_x^4 u + \varepsilon\partial_y^2 u = 0 \ \text{ in } \ C([0,T], H^{s-4}).$$

Since $\partial_t\partial_x u = \partial_x\partial_t u$ in $\mathcal{D}'((0,T)\times\mathbb{R}^2)$, u solves (1.1) at least in the distribution sense.

Remark 2. Note that the best known result for the local well-posedness of KP II goes down to $H^{s_1,0}$, $s_1 > -1/3$ [16]. In view of this result, one would expect the Cauchy problem for the KPB II equation also to be well posed in $H^{s_1,0}$, $s_1 > -1/3$ and perhaps below (we do not consider this question in this paper). For instance, in the case of the KdV equation, the best known result for the local well-posedness goes down to H^s, $s > -3/4$ (see [9]), while in [13] we prove (by the approach developed in this paper) that the Cauchy problem for the KdV Burgers equation is well posed below $H^{-3/4}$.

Remark 3. Another natural way to extend the Bourgain's method would be to consider the norm $\|f\|_{c,s_1,s_2} = \|\langle\xi\rangle^{s_1}\langle\eta\rangle^{s_2}\langle\tau - P(\xi,\eta) + i\xi^2\rangle^c\hat{f}\|_{L^2_{t,z}}$, which is in fact equivalent to $\|f\|_{X^{c,s_1,s_2}} + \|f\|_{L^2_t H^{s_1+2c,s_2}}$. Although the linear estimate on the free term for this norm is the same as the one in (2.5), it seems that the estimate on the forcing term (in this norm) leads to the loss of $2c$ x-derivatives in comparison with the one derived in (2.35). This is why we have used the norm defined by (1.4) rather than this one.

This paper is organized as follows: In section 2, we derive estimates in Bourgain spaces on the linear operators W and L. This process is quite general and can be adapted to other dissipative dispersive semigroups; see [13]. In section 3, we recall some Strichartz's estimates for the KP equations and we use them in the framework of Bourgain spaces. In section 4, we prove a nonlinear estimate which enables us to obtain the local part of Theorem 1.1. Next, we derive a priori estimates to prove the global existence result. Finally, section 5 is devoted to the proof of Theorem 1.2.

2. Linear estimates. In this section we study the two linear operators related to the integral equation (1.5). The following lemmas will be of constant use in the first part of this section.

LEMMA 2.1. *Let s be in $\mathbb{R}$ and $\lambda > 0$.*

(a) *For all $f \in H^s$ we have*

$$\|f(\lambda t)\|_{H^s} \le C\,(\lambda^{-1/2} + \lambda^{s-1/2})\|f(t)\|_{H^s}. \tag{2.1}$$

(b) *For all $f \in \dot{H}^s$ we have*

$$\|f(\lambda t)\|_{\dot{H}^s} \le C\,\lambda^{s-1/2}\|f(t)\|_{\dot{H}^s}. \tag{2.2}$$

LEMMA 2.2. *For all $s \ge 0$, $H^s(\mathbb{R}) \cap L^\infty(\mathbb{R})$ is an algebra and, furthermore,*

$$\|uv\|_{H^s} \le \|u\|_{H^s}\|v\|_{L^\infty} + \|v\|_{\dot{H}^s}\|u\|_{L^\infty}. \tag{2.3}$$

Also, recall that the Fourier transform of $f : t \to e^{-|t||\xi|^2}$ is

$$\mathcal{F}_t(e^{-|t||\xi|^2})(\tau) = \frac{2|\xi|^2}{|\tau|^2 + |\xi|^4}. \tag{2.4}$$

2.1. Linear estimate for the free term.

Proposition 2.3. *Let s, s_1, s_2 be in $\mathbb{R}$ and $b \geq 1/2$.*

(a) *For all $\varphi \in H^{s_1,s_2}$, we have*

$$\|\psi(t)W(t)\varphi\|_{X^{b,s_1,s_2}} \leq C\|\varphi\|_{H^{s_1+2b-1,s_2}}. \tag{2.5}$$

(b) *For all $\varphi \in H^{s}$, we have*

$$\|\psi(t)W(t)\varphi\|_{X^{b,s}} \leq C\|\Delta_x^{\frac{2b-1}{2}}\varphi\|_{H^s}. \tag{2.6}$$

Remark. To avoid the loss of x-derivative in the linear estimates (2.5) and (2.6), we have to choose $b = 1/2$. It will be therefore natural to solve (1.1) in the spaces $X^{1/2,s}$ and $X^{1/2,s_1,s_2}$.

Proof. By definition,

$$\begin{aligned}\|\psi(t)W(t)\varphi\|_{X^{b,s_1,s_2}} &= \left\|\langle\xi\rangle^{s_1}\langle\eta\rangle^{s_2}\langle\tau - P(\zeta)\rangle^b\mathcal{F}_t(\psi(t)e^{-|t||\xi|^2}e^{itP(\zeta)}\hat{\varphi}(\zeta))(\tau)\right\|_{L^2(\mathbb{R}^3)}\\ &= \left\|\langle\xi\rangle^{s_1}\langle\eta\rangle^{s_2}\hat{\varphi}(\zeta)\ \left\|\langle\tau\rangle^b\mathcal{F}_t(\psi(t)e^{-|t||\xi|^2})\right\|_{L^2(\mathbb{R})}\right\|_{L^2(\mathbb{R}^2)}. \end{aligned}\tag{2.7}$$

Now we estimate g defined by

$$g = \left\|\langle\tau\rangle^b\mathcal{F}_t(\psi(t)e^{-|t||\xi|^2})\right\|_{L^2} = \left\|\psi(t)e^{-|t||\xi|^2}\right\|_{H^b}.$$

By virtue of Lemma 2.2 we obtain

$$g \leq \|\psi\|_{H^b}\|e^{-|t||\xi|^2}\|_{L^\infty} + \|\psi\|_{L^\infty}\|e^{-|t||\xi|^2}\|_{\dot{H}^b},$$

and by Lemma 2.1,

$$g \leq C(1 + |\xi|^{2b-1}). \tag{2.8}$$

Hence, since $b \geq 1/2$, gathering (2.7) and (2.8) we obtain

$$\|\psi(t)W(t)\varphi\|_{X^{b,s_1+2b-1,s_2}} \leq C\|\langle\xi\rangle^{s_1+2b-1}\langle\eta\rangle^{s_2}\hat{\varphi}(\zeta)\|_{L^2},$$

which ends the proof of (2.5). The proof of (2.6) is similar. □

2.2. Linear estimates for the forcing term. We study first some smoothing properties between one dimensional Sobolev spaces for the linear operator

$$K : f \mapsto \mathbb{1}_{\mathbb{R}_+}(t)\psi(t)\int_0^t e^{-|t-t'||\xi|^2}f(t')\,dt'. \tag{2.9}$$

Next, we will use these results to obtain some smoothing properties between X^{b,s_1,s_2} and $X^{b,s}$ spaces for the linear operator

$$L : f \mapsto \mathbb{1}_{\mathbb{R}_+}(t)\psi(t)\int_0^t W(t-t')f(t')\,dt'. \tag{2.10}$$

2.2.1. Linear estimates for K. In this section we study the boundedness properties of K. Our main result is the following.

PROPOSITION 2.4. *Let $0 \leq b < 1/2$. For f in $H^{-b}(\mathbb{R})$ consider g defined on $\mathbb{R}$ by*

$$g(t) = \mathbb{1}_{\mathbb{R}_+}(t)\psi(t)\int_0^t e^{-|t-t'||\xi|^2} f(t')\,dt'. \tag{2.11}$$

Then it holds that

$$\forall \xi \in \mathbb{R},\ \|g\|_{H^{1/2}(\mathbb{R})} \leq C\langle\xi\rangle^{-(1-2b)}\|f\|_{H^{-b}}. \tag{2.12}$$

Proof. By a straightforward calculation we have

$$\begin{aligned} g(t) &= \mathbb{1}_{\mathbb{R}_+}(t)\psi(t)e^{-|t||\xi|^2}\int_0^t e^{t'|\xi|^2} f(t')\,dt' \\ &= \mathbb{1}_{\mathbb{R}_+}(t)\psi(t)e^{-|t||\xi|^2}\int_0^t e^{t'|\xi|^2}\int_{\mathbb{R}} e^{it'\tau}\hat{f}(\tau)\,d\tau\,dt' \\ &= \mathbb{1}_{\mathbb{R}_+}(t)\psi(t)e^{-|t||\xi|^2}\int_{\mathbb{R}} \hat{f}(\tau)\int_0^t e^{t'|\xi|^2}e^{it'\tau}\,dt'\,d\tau \\ &= \mathbb{1}_{\mathbb{R}_+}(t)\psi(t)\int_{\mathbb{R}} \frac{e^{it\tau} - e^{-|\xi|^2|t|}}{i\tau + |\xi|^2}\hat{f}(\tau)\,d\tau. \end{aligned} \tag{2.13}$$

Let us now consider the function k defined on $\mathbb{R}$ by

$$k(t) = \psi(t)\int_{\mathbb{R}} \frac{e^{it\tau} - e^{-|\xi|^2|t|}}{i\tau + |\xi|^2}\hat{f}(\tau)\,d\tau.$$

Since $g(0) = k(0) = 0$ and also

$$\forall t \in \mathbb{R}_+\,,\ k(t) = g(t) \text{ and } \forall t \in \mathbb{R}_-\,,\ g(t) = 0,$$

we have $\|g\|_{L^2} \leq \|k\|_{L^2}$, $\|g\|_{H^1} \leq \|k\|_{H^1}$, and $\|g\|_{H^{1/2}} \leq \|k\|_{H^{1/2}}$. Hence it is enough to prove that

$$\|k\|_{H^{1/2}} \leq C\langle\xi\rangle^{-(1-2b)}\|f\|_{H^{-b}}.$$

To do this we split k into $k = k_1 + k_2$, where

$$k_1(t) = \psi(t)\int_{\mathbb{R}} \frac{1 - e^{-|\xi|^2|t|}}{i\tau + |\xi|^2}\hat{f}(\tau)\,d\tau \tag{2.14}$$

and

$$k_2(t) = \psi(t)\int_{\mathbb{R}} \frac{e^{it\tau} - 1}{i\tau + |\xi|^2}\hat{f}(\tau)\,d\tau. \tag{2.15}$$

(a) Estimate of $\|k_1\|_{H^{1/2}}$.

(a1) We first consider the case $|\xi| \geq 1$. We then have

$$\begin{aligned}\|k_1\|_{H^{1/2}} &\leq \|\psi(t)(1-e^{-|\xi|^2|t|})\|_{H^{1/2}} \left|\int_{\mathbb{R}} \frac{\hat{f}(\tau)\,d\tau}{i\tau+|\xi|^2}\right| \\ &\leq \Big(\|\psi\|_{H^{1/2}} + \|\psi(t)e^{-|t||\xi|^2}\|_{H^{1/2}}\Big) \left|\int_{\mathbb{R}} \frac{\hat{f}(\tau)\,d\tau}{i\tau+|\xi|^2}\right| \\ &\leq C\Big(1+\|\psi(t)\|_{H^{1/2}}\|e^{-|\xi|^2|t|}\|_{L^\infty} + \|\psi(t)\|_{L^\infty}\|e^{-|\xi|^2|t|}\|_{\dot{H}^{1/2}}\Big) \left|\int_{\mathbb{R}} \frac{\hat{f}(\tau)\,d\tau}{i\tau+|\xi|^2}\right|\end{aligned}$$

by virtue of Lemma 2.2. Now applying the Cauchy–Schwarz inequality and Lemma 2.1 we obtain

$$\|k_1\|_{H^{1/2}} \leq C\|f\|_{H^{-b}} \left(\int_{\mathbb{R}} \frac{\langle\tau\rangle^{2b}\,d\tau}{\tau^2+|\xi|^4}\right)^{1/2} \leq C(1+|\xi|^{2b})|\xi|^{-1}\|f\|_{H^{-b}}.$$

Hence

$$\forall\xi,\ |\xi|\geq 1,\ \|k_1\|_{H^{1/2}} \leq C\langle\xi\rangle^{2b-1}\|f\|_{H^{-b}}. \tag{2.16}$$

(a2) We now consider the case $|\xi| \leq 1$. As previously, by the Cauchy–Schwarz inequality

$$\begin{aligned}\|k_1\|_{H^{1/2}} &\leq \|\psi(t)(1-e^{-|\xi|^2|t|})\|_{H^{1/2}} \left|\int_{\mathbb{R}} \frac{\hat{f}(\tau)\,d\tau}{i\tau+|\xi|^2}\right| \\ &\leq \|\psi(t)(1-e^{-|\xi|^2|t|})\|_{H^{1/2}}(1+|\xi|^{2b})|\xi|^{-1}\|f\|_{H^{-b}},\end{aligned}$$

and so, since $|\xi| \leq 1$,

$$\|k_1\|_{H^{1/2}} \leq C\|\psi(t)(1-e^{-|\xi|^2|t|})\|_{H^{1/2}}|\xi|^{-1}\|f\|_{H^{-b}}. \tag{2.17}$$

Now, following [10],

$$\|\psi(t)(1-e^{-|\xi|^2|t|})\|_{H^{1/2}} \leq \sum_{n\geq 1}\left\|\frac{t^n\psi(t)|\xi|^{2n}}{n!}\right\|_{H^{1/2}} \leq C\sum_{n\geq 1}\frac{|\xi|^{2n}}{n!}\|t^n\psi(t)\|_{H^{1/2}}.$$

Note that

$$\|t^n\psi(t)\|_{H^{1/2}} \leq C\|t^n\psi(t)\|_{H^1} \leq Cn,$$

and since $|\xi| \leq 1$, we get

$$\forall\xi,\ |\xi|\leq 1,\ \|\psi(t)(1-e^{-|\xi|^2|t|})\|_{H^{1/2}} \leq C|\xi|^2; \tag{2.18}$$

then, gathering (2.17) and (2.18) we obtain

$$\forall\xi,\ |\xi|\leq 1,\ \|k_1\|_{H^{1/2}} \leq C\langle\xi\rangle^{2b-1}\|f\|_{H^{-b}}. \tag{2.19}$$

Now, (2.16) together with (2.19) prove that

$$\forall \xi \in \mathbb{R},\ \|k_1\|_{H^{1/2}} \leq C\langle \xi\rangle^{2b-1}\|f\|_{H^{-b}}. \tag{2.20}$$

(b) Estimate of $\|k_2\|_{H^{1/2}}$. We split k_2 into $k_2 = k_{2,0} + k_{2,\infty}$, where

$$k_{2,0} = \psi(t)\int_{|\tau|\leq 1} \frac{e^{it\tau}-1}{i\tau+|\xi|^2}\hat{f}(\tau)\,d\tau \tag{2.21}$$

and

$$k_{2,\infty} = \psi(t)\int_{|\tau|\geq 1} \frac{e^{it\tau}-1}{i\tau+|\xi|^2}\hat{f}(\tau)\,d\tau. \tag{2.22}$$

(c) Estimate of $\|k_{2,0}\|_{H^{1/2}}$.

(c1) We first consider the case $|\xi| \geq 1$. By definition

$$\begin{aligned}\|k_{2,0}\|_{H^{1/2}} &= \left\|\psi(t)\int_{|\tau|\leq 1}\sum_{n\geq 1}\frac{(it\tau)^n}{n!(i\tau+|\xi|^2)}\hat{f}(\tau)\,d\tau\right\|_{H^{1/2}}\\ &\leq \sum_{n\geq 1}\frac{\|t^n\psi(t)\|_{H^{1/2}}}{n!}\int_{|\tau|\leq 1}\left|\frac{(i\tau)^n}{i\tau+|\xi|^2}\hat{f}(\tau)\right|\,d\tau\\ &\leq \sum_{n\geq 1}\frac{\|t^n\psi(t)\|_{H^{1/2}}}{n!}\int_{|\tau|\leq 1}\frac{|\hat{f}(\tau)|}{\sqrt{|\tau|^2+|\xi|^4}}\,d\tau.\end{aligned}$$

In the same way as previously and by the Cauchy–Schwarz inequality we get

$$\|k_{2,0}\|_{H^{1/2}} \leq C\left(\sum_{n\geq 1}\frac{1}{(n-1)!}\right)\|f\|_{H^{-b}}\left(\int_{\mathbb{R}}\frac{\langle\tau\rangle^{2b}}{\tau^2+|\xi|^4}\,d\tau\right)^{1/2},$$

and so,

$$\forall \xi,\ |\xi|\geq 1,\ \|k_{2,0}\|_{H^{1/2}} \leq C\langle\xi\rangle^{2b-1}\|f\|_{H^{-b}}. \tag{2.23}$$

(c2) We now consider the case $|\xi| \leq 1$. From the previous calculations,

$$\begin{aligned}\|k_{2,0}\|_{H^{1/2}} &\leq C\left(\sum_{n\geq 1}\frac{\|t^n\psi(t)\|_{H^{1/2}}}{n!}\int_{|\tau|\leq 1}\left|\frac{\tau^n\hat{f}(\tau)}{i\tau+|\xi|^2}\right|\,d\tau\right)\\ &\leq C\|f\|_{H^{-b}}\left(\sum_{n\geq 1}\frac{1}{(n-1)!}\left(\int_{|\tau|\leq 1}\frac{|\tau|^{2n}\langle\tau\rangle^{2b}}{|\tau|^2+|\xi|^4}\,d\tau\right)^{1/2}\right).\end{aligned}$$

Furthermore, for all ξ and for all $n \geq 1$,

$$\left(\int_{|\tau|\leq 1}\frac{|\tau|^{2n}\langle\tau\rangle^{2b}}{|\tau|^2+|\xi|^4}\,d\tau\right)^{1/2} \leq \left(\int_{|\tau|\leq 1}|\tau|^{2(n-1)}\langle\tau\rangle^{2b}\,d\tau\right)^{1/2} \leq C,$$

and so it follows that

$$\forall \xi,\ |\xi| \le 1,\ \|k_{2,0}\|_{H^{1/2}} \le C\|f\|_{H^{-b}}. \tag{2.24}$$

Hence, for $b < 1/2$, gathering (2.23) and (2.24) we obtain

$$\forall \xi \in \mathbb{R},\ \|k_{2,0}\|_{H^{1/2}} \le C\langle \xi\rangle^{2b-1}\|f\|_{H^{-b}}. \tag{2.25}$$

(d) Estimate of $\| k_{2,\infty}\|_{H^{1/2}}$.

(d1) First assume that $|\xi| \le 1$. Then we have

$$\|k_{2,\infty}\|_{H^{1/2}} \le I + J,$$

where

$$I = \left\| \psi(t) \int_{|\tau|\ge 1} \frac{\hat{f}(\tau)}{i\tau + |\xi|^2}\, d\tau \right\|_{H^{1/2}}, \tag{2.26}$$

and

$$J = \left\| \psi(t) \int_{|\tau|\ge 1} \frac{e^{it\tau}\hat{f}(\tau)}{i\tau + |\xi|^2}\, d\tau \right\|_{H^{1/2}}. \tag{2.27}$$

By the Cauchy–Schwarz inequality we get

$$I \le \|\psi\|_{H^{1/2}} \|f\|_{H^{-b}} \left(\int_{|\tau|\ge 1} \frac{\langle \tau\rangle^{2b}}{|\tau|^2}\, d\tau \right)^{1/2},$$

and since $b < 1/2$,

$$\forall \xi,\ |\xi| \le 1,\ I \le C\|f\|_{H^{-b}}. \tag{2.28}$$

To estimate J for $|\xi| \le 1$, observe that

$$J = \left\| \psi(t)\mathcal{F}_t \left(\frac{\hat{f}(\tau)\mathbb{1}_{|\tau|\ge 1}}{i\tau + |\xi|^2} \right)(t) \right\|_{H^{1/2}},$$

and since $\|\mathcal{F}(u)\|_{L^\infty} \le \|u\|_{L^1}$, Lemma 2.2 proves that

$$\begin{aligned} J &\le C \left\| \frac{\hat{f}(\tau)\mathbb{1}_{|\tau|\ge 1}}{i\tau + |\xi|^2} \right\|_{L^1} + C \left\| \mathcal{F}_t \left(\frac{\hat{f}(\tau)\mathbb{1}_{|\tau|\ge 1}}{i\tau + |\xi|^2} \right) \right\|_{\dot{H}^{1/2}} \\ &\le C\|f\|_{H^{-b}} \left[\left(\int_{|\tau|\ge 1} \frac{\langle \tau\rangle^{2b} d\tau}{|\tau|^2 + |\xi|^4} \right)^{1/2} + \sup_{|\tau|\ge 1} \left(\frac{|\tau|\langle \tau\rangle^{2b}}{|\tau|^2 + |\xi|^4} \right)^{1/2} \right] \\ &\le C\|f\|_{H^{-b}} \left[\left(\int_{|\tau|\ge 1} \frac{\langle \tau\rangle^{2b} d\tau}{|\tau|^2} \right)^{1/2} + \sup_{|\tau|\ge 1} \left(\frac{\langle \tau\rangle^{1+2b}}{|\tau|^2} \right)^{1/2} \right] \\ &\le C\|f\|_{H^{-b}}. \end{aligned} \tag{2.29}$$

Next, gathering (2.28) and (2.29), we obtain

$$\forall\xi,\ |\xi|\le 1,\ \|k_{2,\infty}\|_{H^{1/2}} \le C\|f\|_{H^{-b}}. \tag{2.30}$$

(d2) Now assume that $|\xi|\ge 1$. Then by the Cauchy–Schwarz inequality,

$$I \le \|\psi\|_{H^{1/2}}\|f\|_{H^{-b}}\left(\int_{|\tau|\ge 1}\frac{\langle\tau\rangle^{2b}}{|\tau|^2+|\xi|^4}\,d\tau\right)^{1/2} \le C\langle\xi\rangle^{2b-1}\|f\|_{H^{-b}}. \tag{2.31}$$

On the other hand, in the same way as for $|\xi|\le 1$, we have

$$J \le C\|f\|_{H^{-b}}\left[\left(\int_{|\tau|\ge 1}\frac{\langle\tau\rangle^{2b}d\tau}{|\tau|^2+|\xi|^4}\right)^{1/2} + \sup_{|\tau|\ge 1}\left(\frac{|\tau|\langle\tau\rangle^{2b}}{|\tau|^2+|\xi|^4}\right)^{1/2}\right],$$

and it follows that

$$\forall\xi,\ |\xi|\ge 1,\ J\le C\langle\xi\rangle^{2b-1}\|f\|_{H^{-b}}. \tag{2.32}$$

Hence from (2.31) and (2.32) we deduce that

$$\forall\xi,\ |\xi|\ge 1,\ \|k_{2,\infty}\|_{H^{1/2}} \le C\langle\xi\rangle^{2b-1}\|f\|_{H^{-b}}, \tag{2.33}$$

which together with (2.30) proves that

$$\forall\xi\in\mathbb{R},\ \|k_{2,\infty}\|_{H^{1/2}} \le C\langle\xi\rangle^{2b-1}\|f\|_{H^{-b}}. \tag{2.34}$$

This ends the proof of Proposition 2.4. □

2.2.2. Linear estimates for *L*. Now, using Proposition 2.4, we prove some smoothing properties in Bourgain spaces for L defined by (2.10).

PROPOSITION 2.5. *Let s, s_1, s_2 be in $\mathbb{R}$ and $0\le b<1/2$.*

(a) *For all $f\in\mathcal{S}'$ we have*

$$\left\|\mathbb{1}_{\mathbb{R}_+}(t)\psi(t)\int_0^t W(t-t')f(t')\,dt'\right\|_{X^{1/2,s_1,s_2}} \le C\|f\|_{X^{-b,s_1+2b-1,s_2}}. \tag{2.35}$$

(b) *For all $f\in\mathcal{S}'$ we have*

$$\left\|\mathbb{1}_{\mathbb{R}_+}(t)\psi(t)\int_0^t W(t-t')f(t')\,dt'\right\|_{X^{1/2,s}} \le C\|\,\Delta_x^{\frac{2b-1}{2}}f\|_{X^{-b,s}}. \tag{2.36}$$

Proof. We first prove (2.35). By definition

$$\left\|\mathbb{1}_{\mathbb{R}_+}(t)\psi(t)\int_0^t W(t-t')f(t')\,dt'\right\|^2_{X^{1/2,s_1,s_2}}$$

$$= \left\|\langle\xi\rangle^{s_1}\langle\eta\rangle^{s_2}\langle\tau-P(\zeta)\rangle^{\frac12}\mathcal{F}_{t,z}\left(\mathbb{1}_{\mathbb{R}_+}(t)\psi(t)\int_0^t W(t-t')f(t')\,dt'\right)\right\|^2_{L^2(\mathbb{R}^3)}.$$

On the other hand,

$$\mathcal{F}_{t,z}\left(\mathbb{1}_{\mathbb{R}_+}(t)\psi(t)\int_0^t W(t-t')f(t')\,dt'\right)(\tau,\zeta)$$

$$= \mathcal{F}_t \left(\mathbb{1}_{\mathbb{R}_+}(t)\psi(t) \int_0^t e^{-|t-t'||\xi|^2} e^{iP(\zeta)(t-t')} \mathcal{F}_z(f)(t',\zeta)\, dt' \right)(\tau)$$

$$= \mathcal{F}_t \left(\mathbb{1}_{\mathbb{R}_+}(t)\psi(t) \int_0^t e^{-|t-t'||\xi|^2} e^{-iP(\zeta)t'} \mathcal{F}_z(f)(t',\zeta)\, dt' \right)(\tau - P(\zeta)).$$

Hence, performing the change of variable $\tau' = \tau - P(\zeta)$, we obtain

$$\left\| \mathbb{1}_{\mathbb{R}_+}(t)\psi(t) \int_0^t W(t-t') f(t')\, dt' \right\|^2_{X^{1/2,s_1,s_2}}$$

$$= \left\| \mathbb{1}_{\mathbb{R}_+}(t)\psi(t) \int_0^t e^{-|t-t'||\xi|^2} e^{-iP(\zeta)t'} \langle\xi\rangle^{s_1} \langle\eta\rangle^{s_2} \mathcal{F}_x z(f)(t',\zeta)\, dt' \right\|^2_{L^2_\zeta(H^{1/2}_t)}$$

$$= \left\| \mathbb{1}_{\mathbb{R}_+}(t)\psi(t) \int_0^t e^{-|t-t'||\xi|^2} \langle\xi\rangle^{s_1} \langle\eta\rangle^{s_2} \mathcal{F}_z \left(U(-t') f \right)(t',\zeta)\, dt' \right\|^2_{L^2_\zeta(H^{1/2}_t)} .$$

To conclude, we apply Proposition 2.4 to h_ζ defined by

$$h_\zeta(t') = \langle\xi\rangle^{s_1} \langle\eta\rangle^{s_2} \mathcal{F}_z \left(U(-t') f \right)(t',\zeta) \, , \tag{2.37}$$

and we obtain

$$\left\| \mathbb{1}_{\mathbb{R}_+}(t)\psi(t) \int_0^t W(t-t') f(t')\, dt' \right\|^2_{X^{1/2,s_1,s_2}}$$

$$\leq C \int_{\mathbb{R}^2} \langle\xi\rangle^{2(s_1+2b-1)} \langle\eta\rangle^{2s_2} \|\mathcal{F}_z \left(U(-t) f \right)(t,\zeta)\|^2_{H^{-b}}\, d\xi$$

$$\leq C \int_{\mathbb{R}^2} \langle\xi\rangle^{2(s_1+2b-1)} \langle\eta\rangle^{2s_2} \|\langle\tau\rangle^{-b} \hat{f}(\tau + P(\zeta),\zeta)\|^2_{L^2}\, d\xi$$

$$\leq C \|f\|^2_{X^{-b,s_1+2b-1,s_2}} .$$

The proof of (2.36) is the same, up to some obvious modifications. □

As explained previously, it will be convenient to prove local well-posedness of KPB in the space $X^{1/2,s_1,s_2}$ ($s_1 > 0$, $s_2 \geq 0$) and local well-posedness of KPB II in the space $X^{1/2,s}$ ($s \geq 0$). Nevertheless, since the embedding $H^{1/2}(\mathbb{R}) \hookrightarrow L^\infty(\mathbb{R})$ does not hold, we will need Proposition 2.6 below to prove that a solution of (1.1) in $X^{1/2,s_1,s_2}$ (respectively, in $X^{1/2,s}$) belongs also to the space $C([0,T], H^{s_1,s_2})$ (respectively, to $C([0,T], H^s)$).

Proposition 2.6. *Let* $0 \leq b < 1/2$.

(a) *For all* $f \in \mathcal{S}'(\mathbb{R}^3)$ *with* $\Delta_x^b f \in X^{-b,s_1,s_2}$,

$$t \mapsto \int_0^t W(t-t') \partial_x f(t')\, dt' \in C(\mathbb{R}_+, H^{s_1,s_2}). \tag{2.38}$$

Moreover, if (f_n) *is a sequence with* $\Delta_x^b f_n \underset{n\to\infty}{\longrightarrow} 0$ *in* X^{-b,s_1,s_2}, *then*

$$\left\| \int_0^t W(t-t') \partial_x f_n(t')\, dt' \right\|_{L^\infty(\mathbb{R}_+, H^{s_1,s_2})} \underset{n\to\infty}{\longrightarrow} 0. \tag{2.39}$$

(b) *For all $f \in \mathcal{S}'(\mathbb{R}^3)$ with $\Delta_x^b f \in X^{-b,s}$,*

$$t \mapsto \int_0^t W(t-t')\partial_x f(t')\,dt' \in C(\mathbb{R}_+, H^s). \tag{2.40}$$

Moreover, if (f_n) is a sequence with $\Delta_x^b f_n \underset{n\to\infty}{\longrightarrow} 0$ in $X^{-b,s}$, then

$$\left\| \int_0^t W(t-t')\partial_x f_n(t')\,dt' \right\|_{L^\infty(\mathbb{R}_+, H^s)} \underset{n\to\infty}{\longrightarrow} 0. \tag{2.41}$$

Proof. Without loss of generality we can set $s_1 = s_2 = 0$. As noticed in [8], since $U(\cdot)$ is a strongly continuous unitary group in $L^2(\mathbb{R}^2)$, it is enough to prove that

$$F \ : \ t \mapsto U(-t) \int_0^t W(t-t')\partial_x f(t')\,dt'$$

is continuous from $\mathbb{R}_+$ to $L^2(\mathbb{R}^2)$. Setting

$$g(t) = \mathcal{F}_x\Big(U(-t)\partial_x f(t)\Big), \qquad t \in \mathbb{R},$$

(2.38) will thus be proven if we show the continuity of

$$G \ : \ t \mapsto \int_0^t e^{-|\xi|^2(t-t')}\, g(t', \zeta)\,dt' \tag{2.42}$$

for $\langle\tau\rangle^{-b}\langle\xi\rangle^{(2b-1)}\mathcal{F}_t(g) \in L^2_{\tau,\zeta}(\mathbb{R}^3)$. Applying the Fubini theorem, one infers that

$$\begin{aligned} G(t) &= e^{-|\xi|^2 t} \int_{\mathbb{R}} \hat{g}(\tau, \zeta) \int_0^t e^{(|\xi|^2 + i\tau)\,t'}\,dt'\,d\tau \\ &= \int_{\mathbb{R}} \hat{g}(\tau, \zeta)\, \frac{e^{it\tau} - e^{-|\xi|^2 t}}{|\xi|^2 + i\tau}\,d\tau. \end{aligned}$$

Hence,

$$G(t_1) - G(t_2) = \int_{\mathbb{R}} \frac{\hat{g}(\tau, \zeta)}{|\xi|^2 + i\tau}[(e^{i\tau t_1} - e^{i\tau t_2}) - (e^{-|\xi|^2 t_1} - e^{-|\xi|^2 t_2})]\,d\tau. \tag{2.43}$$

When $|\xi| \geq 1$, we notice that

$$\begin{aligned} |G(t_1) - G(t_2)| &\leq C \int_{\mathbb{R}} \frac{|\hat{g}(\tau)|}{|\xi|^2 + |\tau|}\,d\tau \\ &\leq C\, \|g\|_{H^{-b}} \left(\int_{\mathbb{R}} \frac{\langle\tau\rangle^{2b}}{|\xi|^4 + |\tau|^2}\,d\tau \right)^{\frac{1}{2}} \\ &\leq C\, \|g\|_{H^{-b}}\, \langle\xi\rangle^{(2b-1)}. \end{aligned} \tag{2.44}$$

Assume now that $|\xi| \leq 1$ and that $|t_1 - t_2| \leq 1$. In this case we first estimate

$$\left| \int_{\mathbb{R}} \quad \frac{\hat{g}(\tau)}{|\xi|^2 + i\tau}(e^{i\tau t_1} - e^{i\tau t_2})\,d\tau \right|$$

$$
\begin{aligned}
&\le |t_1 - t_2| \int_{|\tau|\le 1} \frac{|\hat{g}(\tau)|\,|\tau|}{\sqrt{|\xi|^4+|\tau|^2}}\, d\tau + 2\int_{|\tau|\ge 1} \frac{|\hat{g}(\tau)|}{\sqrt{|\xi|^4+|\tau|^2}}\, d\tau \\
&\le C\,\|g\|_{H^{-b}} \left[\left(\int_{|\tau|\le 1} \frac{|\tau|^2 \langle \tau\rangle^{2b}\, d\tau}{|\xi|^4+|\tau|^2} \right)^{\frac{1}{2}} + \left(\int_{|\tau|\ge 1} \frac{\langle \tau\rangle^{2b}\, d\tau}{|\xi|^4+|\tau|^2} \right)^{\frac{1}{2}} \right] \\
&\le C\,\|g\|_{H^{-b}} \left[\left(\int_{|\tau|\le 1} \langle \tau\rangle^{2b}\, d\tau \right)^{\frac{1}{2}} + \left(\int_{|\tau|\ge 1} \frac{\langle \tau\rangle^{2b}}{|\tau|^2}\, d\tau \right)^{\frac{1}{2}} \right] \\
&\le C\,\|g\|_{H^{-b}}. \qquad (2.45)
\end{aligned}
$$

It then remains to estimate

$$
\begin{aligned}
\left| \int_{\mathbb{R}} \frac{\hat{g}(\tau)}{|\xi|^2+i\tau} (e^{-|\xi|^2 t_1} - e^{-|\xi|^2 t_2})\, d\tau \right| &\le |t_1-t_2|\, \|g\|_{H^{-b}}\, |\xi|^2 \left(\int_{\mathbb{R}} \frac{\langle\tau\rangle^{2b}}{|\xi|^4+|\tau|^2}\, d\tau \right)^{\frac{1}{2}} \\
&\le C|t_1-t_2|\, \|g\|_{H^{-b}}\, |\xi|^2 (|\xi|^{-1} + |\xi|^{(2b-1)}) \\
&\le C\|g\|_{H^{-b}}. \qquad (2.46)
\end{aligned}
$$

Therefore, gathering (2.43)–(2.46), one infers that

$$
\|G(t_1)-G(t_2)\|^2_{L^2} \le C \int_{\mathbb{R}^3} \langle\tau\rangle^{-2b} \, \langle\xi\rangle^{2(2b-1)}\, |\mathcal{F}_t(g)|^2\, d\tau\, d\zeta. \qquad (2.47)
$$

It is clear that the integrand in (2.43) tends to 0 pointwise in (ζ,τ) as soon as $|t_1-t_2| \to 0$ and is bounded uniformly in $|t_1-t_2|$ by the integrand of the right member of (2.47). The result follows then from Lebesgue dominated convergence theorem.

To show (2.39), it suffices to notice that one has

$$
\sup_{t\in\mathbb{R}_+} \|G_n(t)\|_{L^2(\mathbb{R}^2)} \le C \int_{\mathbb{R}^3} \langle\tau\rangle^{-2b} \, \langle\xi\rangle^{2(2b-1)}\, |\mathcal{F}_t(g_n)|^2\, d\tau\, d\xi,
$$

where G_n is defined as G with $g_n(\cdot) = \mathcal{F}_x(U(-\cdot)\partial_x f_n(\cdot))$ instead of g. Finally the proof of part (b) is similar. □

3. Strichartz estimates for the KP equation. The aim of this section is to prove Lemma 3.4. It will be useful in the following sections while proving some estimates in X^{b,s_1,s_2} spaces and $X^{b,s}$ spaces for the nonlinear term $\partial_x(u^2)$.

LEMMA 3.1. *Let* $2 \le r$ *and* $0 \le \beta \le 1/2$. *Then*

$$
\forall t \ne 0, \ \left\| |D_x|^{-\beta\delta(r)} U(t)\varphi \right\|_{L^r_z} \le C|t|^{-(1-\beta/3)\delta(r)} \|\varphi\|_{L^{\bar{r}}_z}, \qquad (3.1)
$$

where

$$
\frac{1}{\bar{r}} = 1 - \frac{1}{r}\,, \qquad \delta(r) = 1 - \frac{2}{r}. \qquad (3.2)
$$

Proof. From (1.2), $U(t)\varphi = G(t) * \varphi$, where

$$
G(t,x,y) = \int_{\mathbb{R}^2} e^{it\left(\xi^3 - \varepsilon\eta^2/\xi\right)} e^{i(x\xi+y\eta)}\, d\xi\, d\eta.
$$

Noticing as in [15] that

$$\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{i\varepsilon t\eta^2/\xi} e^{iy\eta}\, d\eta = e^{i\varepsilon\, sgn(\xi/t)\frac{\pi}{4}}\sqrt{\frac{|\xi|}{2|t|}}\, e^{-i\varepsilon\xi y^2/(4t)},$$

we have

$$\|\, |D_x|^{-\beta} G(t,x,y)\|_{L^\infty_z} \le C|t|^{-\frac{1}{2}}\int_{\mathbb{R}} |\xi|^{\frac{1}{2}-\beta} e^{it\xi^3}\, d\xi.$$

Applying Lemma 2.1 of [10], we obtain that for all $\beta\in[0,1/2]$,

$$\|\, |D_x|^{-\beta} G(t,x,y)\|_{L^\infty_z} \le C|t|^{-(1-\beta/3)}.$$

Therefore by the Young inequality,

$$\|\, |D_x|^{-\beta} U(t)\varphi\|_{L^\infty_z} \le C|t|^{-(1-\beta/3)}\|\varphi\|_{L^1_z},$$

and as $U(\cdot)$ is a unitary group in L^2_z, the result follows by interpolation. □

Thanks to Lemma 3.1 we can now derive some Strichartz's estimates for the KP equations.

LEMMA 3.2. *Let $2\le r$ and $0\le\beta\le 1/2$. Then*

$$\left\|\, |D_x|^{-\frac{\beta\delta(r)}{2}} U(t)\varphi\right\|_{L^{q,r}_{t,z}} \le C\|\varphi\|_{L^2_z} \tag{3.3}$$

for all (q,r,β) fulfilling the condition

$$0\le \frac{2}{q} = \left(1-\frac{\beta}{3}\right)\delta(r) < 1\ . \tag{3.4}$$

Proof. Using (3.1) and the so-called TT^* method, it is a rather classical process (see [7] and the references therein, for instance) to obtain (3.3). □

Now, we state a result which describes some relationships between Strichartz's inequalities for the KP equations and X^{b,s_1,s_2} spaces.

LEMMA 3.3. *Let $v\in L^2(\mathbb{R}^3)$ with supp $v\subset\{(t,x,y)/\,|t|\le T\}$ and let $\varepsilon>0$. Then for all (r,β,θ) with*

$$2\le r<+\infty,\quad 0\le\beta\le\frac{1}{2},\quad 0\le\theta\le 1,\quad 0\le\delta(r)<\frac{\theta}{1-\beta/3}, \tag{3.5}$$

there exists $\mu=\mu(\varepsilon)>0$ such that

$$\left\|\mathcal{F}^{-1}_{t,x}\left(|\xi|^{-\frac{\beta\delta(r)}{2}}\langle\tau-P(\zeta)\rangle^{-\frac{\theta}{2}(1+\varepsilon)}|\hat{v}(\tau,\zeta)|\right)\right\|_{L^{q,r}_{t,z}} \le CT^\mu\|v\|_{L^2(\mathbb{R}^3)}, \tag{3.6}$$

where q is defined by

$$\frac{2}{q} = \left(1-\frac{\beta}{3}\right)\delta(r) + (1-\theta). \tag{3.7}$$

Proof. Let $\hat{u}=|\hat{v}|$. Using Lemma 3.2 together with Lemma 3.3 of [6], we see that for all $\varepsilon>0$

$$\left\|\, |D_x|^{-\frac{\beta\delta(r)}{2}} u\right\|_{L^{q,r}_{t,z}} \le C\|u\|_{X^{1/2+\varepsilon/4,0,0}} \tag{3.8}$$

provided that (3.4) holds. Furthermore, by the definition of X^{b,s_1,s_2}, we have

$$\|u\|_{L^2_{t,z}} = \|u\|_{X^{0,0,0}}. \tag{3.9}$$

Hence, for $0 \le \theta \le 1$, by interpolation,

$$\left\| \, |D_x|^{-\frac{\theta\beta\delta(r)}{2}} u \right\|_{L^{q_1,r_1}_{t,z}} \le C\|u\|_{X^{\theta(\frac{1}{2}+\frac{\varepsilon}{4}),0,0}}, \tag{3.10}$$

where

$$\frac{1}{q_1} = \frac{\theta}{q} + \frac{1-\theta}{2}, \qquad \frac{1}{r_1} = \frac{\theta}{r} + \frac{1-\theta}{2}.$$

Since $\delta(r_1) = \theta\delta(r)$, (3.5) follows from (3.4) and, moreover,

$$\frac{2}{q_1} = \left(1 - \frac{\beta}{3}\right)\delta(r_1) + (1-\theta).$$

Next, using the assumption on the support of u and the results in [8], we get

$$\left\| \, |D_x|^{-\frac{\theta\beta\delta(r)}{2}} u \right\|_{L^{q_1,r_1}_{t,z}} \le CT^{\mu} \, \|u\|_{X^{\theta(\frac{1}{2}+\frac{\varepsilon}{2}),0,0}},$$

which can be rewritten as

$$\left\| \mathcal{F}^{-1}_{t,z}\left(|\xi|^{-\frac{\beta\delta(r_1)}{2}} \hat{u} \right) \right\|_{L^{q_1,r_1}_{t,z}} \le CT^{\mu} \left\| \langle \tau - P(\zeta) \rangle^{\theta(\frac{1}{2}+\frac{\varepsilon}{2})} \hat{u} \right\|_{L^2}.$$

This clearly completes the proof. □

Now, using Lemma 3.3, we state a result which will allow us to obtain some nonlinear estimates in X^{b,s_1,s_2} and $X^{b,s}$ spaces in the next sections.

LEMMA 3.4. *Let f and g be with compact support in $\{(x,y,t) \; / \; |t| \le T\}$. For $b > 0$ small enough, there exists $\mu > 0$ such that for all $h \in L^2(\mathbb{R}^3)$,*

$$\int_{\mathbb{R}^6} \frac{|f(\tau',\zeta')||g(\tau-\tau',\zeta-\zeta')||h(\tau,\zeta)|}{\langle\sigma_1\rangle^{1/2}|\xi'|^b\langle\sigma_2\rangle^{1/2}} \, d\tau d\tau' d\zeta d\zeta'$$

$$\le CT^{\mu}\|f\|_{L^2}\|g\|_{L^2}\|h\|_{L^2} \tag{3.11}$$

and

$$\int_{\mathbb{R}^6} \frac{|f(\tau',\zeta')||g(\tau-\tau',\zeta-\zeta')||h(\tau,\zeta)|}{\langle\sigma_1\rangle^{1/2-b}|\xi'|^b\langle\sigma_2\rangle^{1/2}\langle\sigma\rangle^b} \, d\tau d\tau' d\zeta d\zeta'$$

$$\le CT^{\mu}\|f\|_{L^2}\|g\|_{L^2}\|h\|_{L^2}, \tag{3.12}$$

where σ, σ_1, and σ_2 are defined by

$$\sigma = \tau - P(\zeta), \quad \sigma_1 = \tau' - P(\zeta'), \quad \sigma_2 = \tau - \tau' - P(\zeta - \zeta').$$

Proof. We first prove (3.11). By the Plancherel theorem, the Cauchy–Schwarz inequality, and then by the Hölder inequality, we see that the right-hand side of (3.11) is bounded by

$$\left\| \mathcal{F}^{-1}_{t,z}\left(\frac{|f(\tau,\zeta)|}{\langle\sigma\rangle^{1/2}|\xi|^b} \right) \right\|_{L^{4,r_1}_{t,z}} \left\| \mathcal{F}^{-1}_{t,z}\left(\frac{|g(\tau,\zeta)|}{\langle\sigma\rangle^{1/2}} \right) \right\|_{L^{4,r_2}_{t,z}} \|h\|_{L^2_{t,z}}, \tag{3.13}$$

provided that

$$\frac{1}{r_1}+\frac{1}{r_2}=\frac{1}{2}\,. \tag{3.14}$$

First we apply Lemma 3.3 with $(\beta_2,\theta_2,q_2)=(0,(1+\varepsilon_2)^{-1},4)$ for ε_2 small enough. From (3.7), it follows that $\delta(r_2)=1/2^-$, $r_2=4^-$, and that (3.5) holds. Hence, by virtue of Lemma 3.3 we get

$$\left\|\mathcal{F}^{-1}_{t,z}\left(\frac{|g(\tau,\zeta)|}{\langle\tau-P(\zeta)\rangle^{1/2}}\right)\right\|_{L^{4,r_2}_{t,z}}\le CT^{\mu}\|g\|_{L^2_{t,z}}. \tag{3.15}$$

Now since $r_2=4^-$, in order to fulfill (3.14) we choose

$$r_1=\frac{2r_2}{r_2-2}=4^+,$$

which implies $\delta(r_1)=1/2^+$. Let

$$q_1=4,\quad \theta_1=\frac{1}{1+\varepsilon_1}=1^-,\quad \beta_1=\frac{2}{\delta(r_1)}\,b=4^+\,b.$$

Note that for $b<1/8$, $0\le\beta_1\le 1/2$, and so (3.5) is fulfilled since we have

$$\delta(r_1)\sim\frac{1}{2}<1^-=\theta_1\le\frac{\theta_1}{1-\beta_1/3}.$$

Moreover, for b small enough, we can always find $\varepsilon_1=\varepsilon_1(b)$ such that (3.7) holds. By virtue of Lemma 3.3 it follows that

$$\left\|\mathcal{F}^{-1}_{t,z}\left(\frac{|f(\tau,\zeta)|}{\langle\tau-P(\zeta)\rangle^{1/2}|\xi|^{b}}\right)\right\|_{L^{4,r_1}_{t,z}}\le CT^{\mu}\|f\|_{L^2_{t,z}}. \tag{3.16}$$

The proof of (3.11) follows then from (3.13), (3.15), and (3.16).

Now we prove (3.12). By the Plancherel theorem and the Hölder inequality (first in space and then in time) we obtain that the right-hand side of (3.12) is bounded by the product

$$\left\|\mathcal{F}^{-1}_{t,z}\left(\frac{|g(\tau,\zeta)|}{\langle\sigma\rangle^{1/2}}\right)\right\|_{L^{q_1,r_1}_{t,z}}\left\|\mathcal{F}^{-1}_{t,z}\left(|\xi|^{-b}\frac{|f(\tau,\zeta)|}{\langle\sigma\rangle^{1/2-b}}\right)\right\|_{L^{q_2,r_2}_{t,z}}\left\|\mathcal{F}^{-1}_{t,z}\left(\frac{|h(\tau,\zeta)|}{\langle\sigma\rangle^{b}}\right)\right\|_{L^{q_3,r_3}_{t,z}}, \tag{3.17}$$

provided

$$\frac{1}{q_1}+\frac{1}{q_2}+\frac{1}{q_3}=1,\quad \frac{1}{r_1}+\frac{1}{r_2}+\frac{1}{r_3}=1. \tag{3.18}$$

To apply Lemma 3.3 to each of the three terms in (3.17), for b small enough we take first $\varepsilon_1=\varepsilon_2=\varepsilon_3=\varepsilon$, where $\varepsilon=\varepsilon(b)$ will be a small parameter. Next we take

$$\theta_1=\frac{1}{1+\varepsilon},\quad \theta_2=\frac{1-2b}{1+\varepsilon},\quad \theta_3=\frac{2b}{1+\varepsilon}, \tag{3.19}$$

and we choose $\beta_1=\beta_3=0$. From (3.7), it remains to find β_2, q_i, and r_i with

$$\frac{2}{q_1}=\delta_1+(1-\theta_1),\quad \frac{2}{q_2}=\delta_2+(1-\theta_2)-\frac{\beta_2\delta_2}{3},\quad \frac{2}{q_3}=\delta_3+(1-\theta_3), \tag{3.20}$$

$$\beta_2\delta_2 = 2b, \tag{3.21}$$

such that (3.5) is fulfilled for $i = 1, 2, 3$.

First note that from (3.18) and (3.19), $\sum \delta_i = 1$, $\sum 2/q_i = 2$, and also $\sum \theta_i = 2/(1+\varepsilon)$. Hence, adding the three equations in (3.20), we see that, necessarily, $2 = 4 - 2/(1+\varepsilon) - \beta_2\delta_2/3$; i.e., $\beta_2\delta_2 = 6\varepsilon/(1+\varepsilon)$. This relation is compatible with (3.21) if and only if

$$\frac{6\varepsilon}{1+\varepsilon} = 2b. \tag{3.22}$$

Hence, for b small enough, we choose $\varepsilon(b)$ satisfying (3.22), which defined the values of θ_i through (3.19). Also, from (3.19)–(3.20)

$$\frac{2}{q_1} = \delta_1 + \frac{\varepsilon}{1+\varepsilon}, \quad \frac{2}{q_2} = \delta_2 + \frac{2b-\varepsilon}{1+\varepsilon}, \quad \frac{2}{q_3} = \delta_3 + 1 - \frac{2b}{1+\varepsilon}. \tag{3.23}$$

Now we choose $(r_1, r_2, r_3) = (4^-, 4^-, 2^+)$ such that $\sum r_i^{-1} = 1$. From (3.23) it follows that $\beta_2 = 2b/\delta_2$ and, moreover, (q_1, q_2, q_3) is closed to $(4, 4, 2)$. Also recall that by construction we have $\sum q_i^{-1} = 1$.

It remains to prove that (3.5) is fulfilled for $i = 1, 2, 3$. First remark that, for $b = 0^+$, $\varepsilon(b) = 0^+$. It is clear that $0 \le \theta_i \le 1$, that $0 \le \beta_i \le 1/2$ for $i = 1, 2, 3$, and also that the last restriction in (3.5) is fulfilled for $i = 1, 2$ since $(\delta_1, \delta_2) = (1/2^+, 1/2^+)$ and $(\theta_1, \theta_2) = (1^-, 1^-)$. To see that (3.5) can be also fulfilled for $i = 3$, we remark that it is enough to have $0 \le \delta_3 \le \theta_3$. But since r_3 does not depend on b, this can always be achieved by choosing r_3 close enough to 2.

Therefore from Lemma 3.3 we have

$$\left\| \mathcal{F}^{-1}_{t,z}\left(|\xi|^{-b} \frac{|f(\tau,\zeta)|}{\langle\sigma\rangle^{1/2-b}} \right) \right\|_{L^{q_1,r_1}_{t,z}} \le CT^{\mu} \|f\|_{L^2_{t,z}}, \tag{3.24}$$

$$\left\| \mathcal{F}^{-1}_{t,z}\left(\frac{|g(\tau,\zeta)|}{\langle\sigma\rangle^{1/2}} \right) \right\|_{L^{q_2,r_2}_{t,z}} \le CT^{\mu} \|g\|_{L^2_{t,z}}, \tag{3.25}$$

$$\left\| \mathcal{F}^{-1}_{t,z}\left(\frac{|h(\tau,\zeta)|}{\langle\sigma\rangle^{b}} \right) \right\|_{L^{q_3,r_3}_{t,z}} \le \|h\|_{L^2_{t,z}}. \tag{3.26}$$

Gathering (3.17), (3.24), (3.25), and (3.26) we obtain (3.12). □

4. Nonlinear estimates and applications to the resolution of KPB I.

First we prove existence and uniqueness of a solution u in the space $X^{1/2,s_1,s_2}(\mathbb{R}^2)$. They follow from the linear estimates of Propositions 2.3 and 2.5 together with the nonlinear estimates of Proposition 4.1 below. Next using Proposition 2.6, we prove that u belongs also to $C([0,T], H^{s_1,s_2}(\mathbb{R}^2))$.

Finally, by means of the conservation laws of the classical KP equation, we show a priori estimates which yield the global existence of the solution of KPB I, provided that φ belongs to H^{s_1,s_2}, $(s_1, s_2) \in [1, +\infty[\times\mathbb{R}_+$, with $\mathcal{F}^{-1}_x(\frac{\eta}{\xi}\hat{\varphi}) \in L^2(\mathbb{R}^2)$.

4.1. Nonlinear estimates in X^{b,s_1,s_2}.

PROPOSITION 4.1. *Let $s_1 > 0$ and $s_2 \ge 0$. Let P be defined by*

$$P(\xi,\eta) = \xi^3 + \frac{\eta^2}{\xi} \quad \textit{or} \quad P(\xi,\eta) = \xi^3 - \frac{\eta^2}{\xi}.$$

For u, v with support in the subset $\{(t,x,y)/\,|t|\le T\}$, there exists $\mu>0$ such that the following bilinear estimate holds:

$$\begin{aligned}\|uv\|_{X^{0,s_1,s_2}} \le CT^{\mu}\big(&\|u\|_{X^{1/2,0+,0}}\|v\|_{X^{1/2,s_1,s_2}} + \|u\|_{X^{1/2,0+,s_2}}\|v\|_{X^{1/2,s_1,0}}\\ &+ \|u\|_{X^{1/2,s_1,0}}\|v\|_{X^{1/2,0+,s_2}} + \|u\|_{X^{1/2,s_1,s_2}}\|v\|_{X^{1/2,0+,0}}\big)\,. \end{aligned}\tag{4.1}$$

Proof. By duality it is equivalent to prove that for $\delta>0$ small enough and for all $\omega\in X^{0,-s_1,-s_2}$

$$\begin{aligned}|\langle uv,\omega\rangle| \le CT^{\mu}\Big(&\|u\|_{X^{1/2,\delta,0,}}\|v\|_{X^{1/2,s_1,s_2}} + \|u\|_{X^{1/2,\delta,s_2}}\|v\|_{X^{1/2,s_1,0}}\\ &+ \|u\|_{X^{1/2,s_1,0}}\|v\|_{X^{1/2,\delta,s_2}} + \|u\|_{X^{1/2,s_1,s_2}}\|v\|_{X^{1/2,\delta,0}}\Big)\|\omega\|_{X^{0,-s_1,-s_2}}\,. \end{aligned}\tag{4.2}$$

Now, consider f, g, and h, respectively, defined by

$$\hat{f}(\tau,\zeta) = \langle \tau - P(\zeta)\rangle^{1/2}\langle\xi\rangle^{s_1}\langle\eta\rangle^{s_2}\hat{u}(\tau,\zeta),$$

$$\hat{g}(\tau,\zeta) = \langle \tau - P(\zeta)\rangle^{1/2}\langle\xi\rangle^{s_1}\langle\eta\rangle^{s_2}\hat{v}(\tau,\zeta)\,,\quad \hat{h}(\tau,\zeta) = \frac{\hat{\omega}(\tau,\zeta)}{\langle\xi\rangle^{s_1}\langle\eta\rangle^{s_2}}\,.$$

Since $\|u\|_{X^{1/2,s_1,s_2}} = \|f\|_{L^2_{t,z}}$, $\|v\|_{X^{1/2,s_1,s_2}} = \|g\|_{L^2_{t,z}}$, and $\|\omega\|_{X^{0,-s_1,-s_2}} = \|h\|_{L^2_{t,z}}$, we easily see that (4.2) is equivalent to the inequality

$$\begin{aligned}&\int_{\mathbb{R}^6}\frac{|\hat{f}(\tau-\tau',\zeta-\zeta')||\hat{g}(\tau',\zeta')||\hat{h}(\tau,\zeta)|\langle\xi\rangle^{s_1}\langle\eta\rangle^{s_2}\,d\tau d\zeta d\tau' d\zeta'}{\langle\sigma_2\rangle^{1/2}\langle\sigma_1\rangle^{1/2}\langle\xi-\xi'\rangle^{s_1}\langle\xi'\rangle^{s_1}\langle\eta-\eta'\rangle^{s_2}\langle\eta'\rangle^{s_2}}\\ &\le CT^{\mu}\Big(\|f\|_{L^2_tH_z^{-s_1+\delta,-s_2}}\|g\|_{L^2_{(t,z)}} + \|f\|_{L^2_tH_z^{-s_1+\delta,0}}\|g\|_{L^2_tH_z^{0,-s_2}}\\ &\quad + \|f\|_{L^2_tH_z^{0,-s_2}}\|g\|_{L^2_tH_z^{-s_1+\delta,0}} + \|f\|_{L^2_{(t,z)}}\|g\|_{L^2_tH_z^{-s_1+\delta,-s_2}}\Big)\|h\|_{L^2_{(t,z)}}\,, \end{aligned}\tag{4.3}$$

where σ_1 and σ_2 are defined by

$$\sigma_1 = \tau - P(\zeta)\ \text{ and }\ \sigma_2 = \tau-\tau' - P(\zeta-\zeta').$$

Note that for all $s\ge 0$ we have

$$\frac{\langle\theta\rangle^{s}}{\langle\theta-\theta'\rangle^{s}\langle\theta'\rangle^{s}} \le \frac{C}{\langle\theta-\theta'\rangle^{s}} + \frac{C}{\langle\theta'\rangle^{s}}\,,\ \theta\,,\ \theta'\in\mathbb{R}.$$

It follows that the left-hand side of (4.3) is bounded by the sum $I_1+I_2+I_3+I_4$, where

$$I_1 = \int_{\mathbb{R}^6}\frac{|\hat{f}(\tau-\tau',\zeta-\zeta')||\hat{g}(\tau',\zeta')||\hat{h}(\tau,\zeta)|\,d\tau d\zeta d\tau' d\zeta'}{\langle\sigma_2\rangle^{1/2}\langle\sigma_1\rangle^{1/2}\langle\xi-\xi'\rangle^{s_1}\langle\eta-\eta'\rangle^{s_2}},$$

$$I_2 = \int_{\mathbb{R}^6}\frac{|\hat{f}(\tau-\tau',\zeta-\zeta')||\hat{g}(\tau',\zeta')||\hat{h}(\tau,\zeta)|\,d\tau d\zeta d\tau' d\zeta'}{\langle\sigma_2\rangle^{1/2}\langle\sigma_1\rangle^{1/2}\langle\xi-\xi'\rangle^{s_1}\langle\eta'\rangle^{s_2}},$$

$$I_3 = \int_{\mathbb{R}^6}\frac{|\hat{f}(\tau-\tau',\zeta-\zeta')||\hat{g}(\tau',\zeta')||\hat{h}(\tau,\zeta)|\,d\tau d\zeta d\tau' d\zeta'}{\langle\sigma_2\rangle^{1/2}\langle\sigma_1\rangle^{1/2}\langle\xi'\rangle^{s_1}\langle\eta-\eta'\rangle^{s_2}},$$

$$I_4 = \int_{\mathbb{R}^6} \frac{|\hat{f}(\tau-\tau',\zeta-\zeta')||\hat{g}(\tau',\zeta')||\hat{h}(\tau,\zeta)|\,d\tau d\zeta d\tau' d\zeta'}{\langle\sigma_2\rangle^{1/2}\langle\sigma_1\rangle^{1/2}\langle\xi'\rangle^{s_1}\langle\eta'\rangle^{s_2}}.$$

By Lemma 3.4, one infers that

$$\begin{aligned} I_1 &= \int_{\mathbb{R}^6} \frac{|\hat{f}(\tau',\zeta')|}{\langle\xi'\rangle^{s_1-\delta}\langle\eta'\rangle^{s_2}} \frac{|\hat{g}(\tau-\tau',\zeta-\zeta')||\hat{h}(\tau,\zeta)|}{\langle\sigma_2\rangle^{1/2}\langle\sigma_1\rangle^{1/2}\langle\xi'\rangle^{\delta}}\,d\tau d\zeta d\tau' d\zeta' \\ &\le C\,T^{\mu}\|f\|_{L^2_tH_z^{-s_1+\delta,-s_2}}\|g\|_{L^2_{t,z}}\|h\|_{L^2_{t,z}} \end{aligned}$$

for $\delta > 0$ small enough. In the same way,

$$\begin{aligned} I_2 &\le CT^{\mu}\|f\|_{L^2_tH_z^{-s_1+\delta,0}}\|g\|_{L^2_tH_z^{0,-s_2}}\|h\|_{L^2_{t,z}}, \\ I_3 &\le CT^{\mu}\|f\|_{L^2_tH_z^{0,-s_2}}\|g\|_{L^2_tH_z^{-s_1+\delta,0}}\|h\|_{L^2_{t,z}}, \\ I_4 &\le CT^{\mu}\|f\|_{L^2_{t,z}}\|g\|_{L^2_tH_z^{-s_1+\delta,-s_2}}\|h\|_{L^2_{t,z}}, \end{aligned}$$

which completes the proof of the proposition. □

4.2. Local existence result.

PROPOSITION 4.2. *Let $\varphi \in H^{s_1,s_2}(\mathbb{R}^2)$, $(s_1,s_2) \in \mathbb{R}_+^* \times \mathbb{R}_+$. For any $\delta > 0$, there exist $T = T(\|\varphi\|_{H^{\delta,0}}) > 0$ and a unique local solution u to* (1.1) *in*

$$Y_T = X_T^{1/2,s_1,s_2} \cap C([0,T],H^{s_1,s_2}).$$

Furthermore the map $\varphi \mapsto u$ is continuous from H^{s_1,s_2} to Y_T.

Remark. It is important to note that the lower bound for the time of existence of u in $X_T^{1/2,s_1,s_2} \cap C([0,T],H^{s_1,s_2})$ depends only on the norm of the corresponding initial data in $H^{\delta,0}$ (for all fixed $\delta > 0$).

Proof. Let φ be in H^{s_1,s_2}, $(s_1,s_2) \in \mathbb{R}_+^* \times \mathbb{R}_+$. For $T \le 1$, u is a solution of KPB on $[0,T/2]$ if and only if u is a solution of the integral equation $u = L(u)$ with

$$L(u) = \psi(t)\left[W(t)u_0 - \frac{\mathbb{1}_{\mathbb{R}^+}(t)}{2}\int_0^t W(t-t')\partial_x(\psi_T^2(t')\,u^2(t'))\,dt'\right]. \tag{4.4}$$

We first prove the statement for $T = T(\|\varphi\|_{H^{s_1,0}})$.

Following Bourgain, we are going to solve (4.4) in a ball of the space

$$Z = \{u \in X_T^{1/2,s_1,s_2} \;/\; \|u\|_Z = \|u\|_{X_T^{1/2,s_1,0}} + \gamma\,\|u\|_{X_T^{1/2,s_1,s_2}} < +\infty\},$$

where the constant γ is defined for all nontrivial φ by

$$\gamma = \frac{\|\varphi\|_{H^{s_1,0}}}{\|\varphi\|_{H^{s_1,s_2}}}.$$

From Propositions 2.3 and 2.5 with $b = 0$,

$$\|L(u)\|_{X_T^{1/2,s_1,0}} \le C\|\varphi\|_{s_1,0} + C\|\Delta_x^{-\frac{1}{2}}\partial_x(u^2)\|_{X_T^{0,s_1,0}}.$$

Proposition 4.1 then yields

$$\|L(u)\|_{X_T^{1/2,s_1,0}} \le C\,\|\varphi\|_{H^{s_1,0}} + C\,T^{\mu}\|u\|^2_{X_T^{1/2,s_1,0}}. \tag{4.5}$$

Next, since $\partial_x(u^2) - \partial_x(v^2) = \partial_x[(u-v)(u+v)]$, we get

$$\|L(u) - L(v)\|_{X_T^{1/2,s_1,0}} \le C\,T^{\mu} \|u-v\|_{X_T^{1/2,s_1,0}} \|u+v\|_{X_T^{1/2,s_1,0}}. \tag{4.6}$$

In the same way, we obtain

$$\|L(u)\|_{X_T^{1/2,s_1,s_2}} \le C\,\|\varphi\|_{H^{s_1,s_2}} + CT^{\mu}\|u\|_{X_T^{1/2,s_1,0}}\|u\|_{X_T^{1/2,s_1,s_2}} \tag{4.7}$$

and

$$\begin{aligned}\|L(u) - L(v)\|_{X_T^{1/2,s_1,s_2}} \le C\,T^{\mu}\Big(&\|u-v\|_{X_T^{1/2,s_1,0}}\|u+v\|_{X_T^{1/2,s_1,s_2}} \\ &+ \|u-v\|_{X_T^{1/2,s_1,s_2}}\|u+v\|_{X_T^{1/2,s_1,0}}\Big).\end{aligned} \tag{4.8}$$

Recalling the definition of $\|\cdot\|_Z$, (4.5)–(4.8) lead to

$$\|L(u)\|_Z \le C\,(\|\varphi\|_{H^{s_1,0}} + \gamma\|\varphi\|_{H^{s_1,s_2}}) + C\,T^{\mu}\|u\|_Z^2 \tag{4.9}$$

and

$$\begin{aligned}\|L(u) - L(v)\|_Z \le C\,T^{\mu}\Big(&\|u-v\|_{X_T^{1/2,s_1,0}}\|u+v\|_Z \\ &+ \gamma\|u-v\|_{X_T^{1/2,s_1,s_2}}\|u+v\|_{X_T^{1/2,s_1,0}}\Big) \\ \le{}& C\,T^{\mu}\|u-v\|_Z\|u+v\|_Z.\end{aligned} \tag{4.10}$$

Now, setting $T = (4C^2(\|\varphi\|_{H^{s_1,0}} + \gamma\|\varphi\|_{H^{s_1,s_2}}))^{-1/\mu}$, which yields $T = (8C^2\|\varphi\|_{H^{s_1,0}})^{-1/\mu}$ by definition of γ, we deduce from (4.9) and (4.10) that L is strictly contractive on the ball of radius $2C(\|\varphi\|_{H^{s_1,0}} + \gamma\|\varphi\|_{H^{s_1,s_2}})$ in Z. This proves the existence of a unique solution to (1.1) in $X_T^{1/2,s_1,s_2}$ with $T = T(\|\varphi\|_{H^{s_1,0}}) > 0$.

Note that $\psi(\cdot)W(\cdot)\varphi$ belongs to $C([0,T], H^{s_1,s_2})$ since φ belongs to H^{s_1,s_2}. Moreover, since $u \in X_T^{1/2,s_1,s_2}$, one infers from Proposition 4.1 that $u^2 \in X_T^{0,s_1,s_2}$, and it follows from the first assertion of Proposition 2.6 that

$$t \mapsto \int_0^t W(t-t')\partial_x(u^2(t'))\,dt'$$

also belongs to $C([0,T], H^{s_1,s_2})$. Thus u belongs to $C([0,T], H^{s_1,s_2})$.

Now, standard arguments enable us to extend u on a maximal interval of existence $[0, T_*[$ such that

$$\text{if } \ T_* < +\infty\,, \ \text{then } \limsup_{t \nearrow T_*} \|u(t)\|_{H^{s_1,0}} = +\infty. \tag{4.11}$$

Next, proceeding exactly in the same way as above but in the space

$$\tilde{Z} = \{u \in X_T^{1/2,s_1,0} \ / \ \|u\|_{\tilde{Z}} = \|u\|_{X_T^{1/2,\delta,0}} + \tilde{\gamma}\,\|u\|_{X_T^{1/2,s_1,0}} < +\infty\},$$

where

$$\tilde{\gamma} = \frac{\|\varphi\|_{H^{\delta,0}}}{\|\varphi\|_{H^{s_1,0}}},$$

we obtain that for $\tilde{T} = \tilde{T}(\|\varphi\|_{H^{\delta,0}})$, L is also strictly contractive on a ball of $\tilde{Z}$. Since obviously $H^{s_1,s_2} \subset H^{s_1,0}$, it follows that there exists $\tilde{T} = \tilde{T}(\|\varphi\|_{H^{\delta,0}})$ and a unique solution $\tilde{u}$ to (1.1) in $C([0,\tilde{T}], H^{s_1,0}) \cap X^{1/2,s_1,0}$. By uniqueness, $u = \tilde{u}$ on $[0, \min(\tilde{T}, T_*)[$, and this implies that $T_* \geq \tilde{T}(\|\varphi\|_{H^{\delta,0}})$.

Finally, the continuity of the map $\varphi \mapsto u$ from H^{s_1,s_2} to $X^{1/2,s_1,s_2}$ follows from classical arguments, while the continuity from H^{s_1,s_2} to $C([0,T_*[, H^{s_1,s_2})$ follows from Proposition 2.6. □

4.3. Global existence for KPB I. Let us now show that for $\varepsilon = -1$ and $\varphi \in H^{s_1,s_2}$, (s_1,s_2) in $[1,+\infty[\times\mathbb{R}_+$ with $\mathcal{F}_z^{-1}(\frac{\eta}{\xi}\hat{\varphi}) \in L^2(\mathbb{R}^2)$, the solution to (1.1) is global in time. To establish a priori estimates on the solution, we will use the following lemma which is directly inspired by the conservation laws of the KP equations.

LEMMA 4.3. *Let $u \in C([0,T], H^3(\mathbb{R}^2))$ be a solution to KPB equations ($\varepsilon = \pm 1$) with initial data*

$$\varphi \in \mathcal{N} = \left\{ \phi \in H^3(\mathbb{R}^2),\ \mathcal{F}_z^{-1}\left(\frac{\hat{\phi}}{\xi^2}\right) \in H^3(\mathbb{R}^2) \right\}.$$

Then u satisfies

$$\|u(t)\|_{L^2}^2 + \int_0^t \|u_x\|_{L^2}^2 = \|\varphi\|_{L^2}^2 \quad \forall t \in [0,T] \tag{4.12}$$

and

$$E(u(t)) + \int_0^t \|u_{xx}(\tau)\|_{L^2}^2 + \|u_y(\tau)\|_{L^2}^2 \, d\tau$$

$$= -\frac{1}{2}\int_0^t \int_{\mathbb{R}^2} u^2(\tau) u_{xx}(\tau)\, d\tau + E(\varphi), \tag{4.13}$$

where

$$E(\phi) = \frac{1}{2}\|\phi_x\|_{L^2}^2 - \frac{\varepsilon}{2}\|\partial^{-1}\phi_y\|_{L^2}^2 - \frac{1}{6}\|\phi\|_{L^3}^3.$$

Proof. Proceeding exactly as in [15], one can show that $\partial^{-1}u_{yy}$, $\partial_x^{-1}u_t$, and $\partial^{-2}u_{yy}$ belong to $C([0,T], L^2(\mathbb{R}^2))$. Therefore, we obtain (4.12) (respectively, (4.13)) by applying the operator ∂_x^{-1} on (1.1), multiplying the obtained equation by u (respectively, $-u_{xx} - \varepsilon\partial_x^{-2}u_{yy} - u^2/2$), and integrating by parts in $\mathbb{R}^2$ and then over the time interval $[0,t]$. □

Now, according to Proposition 4.2, denoting by $[0,T_*[$ the maximal interval of existence of the solution u to KPB,

$$\text{if } T_* < +\infty, \quad \text{then } \limsup_{t \nearrow T_*} \|u(t)\|_{H^{1,0}} = +\infty. \tag{4.14}$$

Next, thanks to Lemma 3.2 in [11], one can always find a sequence $(\varphi_n)_{n\geq 0} \subset \mathcal{N} \cap H^{3,3}$ such that

$$\varphi_n \to \varphi \text{ in } \left\{ \phi \in H^{1,0}(\mathbb{R}^2),\ \mathcal{F}_z\left(\frac{\eta}{\xi}\hat{\phi}\right) \in L^2(\mathbb{R}^2) \right\}.$$

Fixing T in $]0,T_*[$, Proposition 4.2 and Lemma 4.3 ensure that for n large enough the solution u_n to KPB with initial data φ_n satisfies (4.12) and (4.13) on $[0,T]$ (note that $H^{3,3} \subset H^3$).

Recalling the anisotropic Sobolev inequality (cf. [1])

$$\|u\|_{L^{2(q+1)}}^{2(q+1)} \leq C\,\|u\|_{L^2}^{2-q}\,\|u_x\|_{L^2}^{2q}\,\|\partial_x^{-1}u_y\|_{L^2}^{q} \quad \forall q \in [0,2], \tag{4.15}$$

we notice that

$$\Big|\int_{\mathbb{R}^2} u^2 u_x\Big| \leq \frac{1}{2}\|u_{xx}\|_{L^2}^2 + \frac{1}{2}\|u\|_{L^2}\|u_x\|_{L^2}^2\|\partial_x^{-1}u_y\|_{L^2}.$$

It then follows from (4.12) and (4.13) that, for all t in $[0,T]$,

$$\begin{aligned} E(u_n(t)) \leq & -\frac{1}{2}\int_0^t \Big(\|u_{n,xx}\|_{L^2}^2 + \|u_{n,y}\|_{L^2}\Big)\,dt \\ & + C\,\|\varphi_n\|_{L^2}^3\|\partial_x^{-1}u_{n,y}\|_{L_T^\infty L_z^2} + E(\varphi_n). \end{aligned} \tag{4.16}$$

Using (4.15) again (with $q = 1/2$) we see that

$$\Big|\int_{\mathbb{R}^2} u_n^3\Big| \leq \|u_n\|_{L^2}^{\frac{3}{2}}\,\|u_{n,x}\|_{L^2}\|\partial_x^{-1}u_{n,y}\|_{L^2}^{\frac{1}{2}},$$

and (4.16) then leads to

$$\begin{aligned} & \|u_{n,x}\|_{L_T^\infty L_z^2}^2 + \|\partial_x^{-1}u_{n,y}\|_{L_T^\infty L_z^2}^2 \\ & \leq C(\|\varphi_n\|_2)\Big(1 + \|u_{n,x}\|_{L_T^\infty L_z^2}^2 + \|\partial_x^{-1}u_{n,y}\|_{L_T^\infty L_z^2}^2\Big)^{\frac{3}{4}} + E(\varphi_n). \end{aligned} \tag{4.17}$$

Since by (4.15), $E(\varphi_n) \leq C\,(\|\varphi_n\|_{H^{1,0}},\, \|\partial_x^{-1}\varphi_{n,y}\|_{L^2})$, it follows from (4.17) that

$$\|u_{n,x}\|_{L_T^\infty L_z^2} \leq C\,\Big(\|\varphi_n\|_{H^{1,0}},\, \|\partial_x^{-1}\varphi_{n,y}\|_{L^2}\Big).$$

By Proposition 4.2, $u_n \to u$ in $C([0,T], H^{1,0})$, and so

$$\|u_x\|_{L_T^\infty L_z^2} \leq C\,\Big(\|\varphi\|_{H^{1,0}},\, \|\partial_x^{-1}\varphi_y\|_{L^2}\Big).$$

Thus (4.14) ensures that $T_* = +\infty$.

5. The Cauchy problem for the KPB II equation in $H^s(\mathbb{R}^2)$, $s \geq 0$. In this section we use the algebraic inequality (5.2) to derive a bilinear estimate which will enable us to show the local existence of a unique solution u in the space $X^{1/2,s}$ for all initial data in the Sobolev space $H^s(\mathbb{R}^2)$, $s \geq 0$. Note that (5.2) holds only when $\varepsilon = +1$ and is crucial to regain the x-derivative in the local well-posedness result for the classical KP II equation. Moreover, we show that the time of existence of the solution in $C([0,T], H^s)$ depends only on the norm of the initial data in $L^2(\mathbb{R}^2)$. Finally, using the decay of the L^2-norm of regular solutions of the KPB II equation, we prove the global existence in $H^s(\mathbb{R}^2)$.

5.1. Nonlinear estimates.

Proposition 5.1. *Let $s \geq 0$ and $b > 0$ small enough. Let $P(\xi, \eta)$ be defined by*

$$P(\xi, \eta) = \xi^3 + \frac{\eta^2}{\xi}.$$

Let u and v supported in the set $\{(t,x,y)/\,|t| \leq T\}$. Then there exists $\mu(b) > 0$ such that the following bilinear estimate holds:

$$\|\Delta_x^{\frac{2b-1}{2}} \partial_x(uv)\|_{X^{-b,s}} \leq CT^{2\mu} \left(\|u\|_{X^{1/2,s}}\|v\|_{X^{1/2,0}} + \|u\|_{X^{1/2,0}}\|v\|_{X^{1/2,s}}\right). \tag{5.1}$$

Proof. By duality it is enough to prove that for all $\omega \in X^{b,-s}$,

$$|\langle \Delta_x^{\frac{2b-1}{2}} \partial_x(uv), \omega\rangle| \leq CT^{2\mu}(\|u\|_{X^{1/2,s}}\|v\|_{X^{1/2,0}} + \|u\|_{X^{1/2,0}}\|v\|_{X^{1/2,s}})\|\omega\|_{X^{b,-s}}.$$

Now, consider

$$\hat{f} = \hat{u}(\tau,\zeta)\langle\zeta\rangle^s\langle\tau - P(\zeta)\rangle^{1/2},\ \hat{g} = \hat{v}(\tau,\zeta)\langle\zeta\rangle^s\langle\tau - P(\zeta)\rangle^{1/2}$$

and

$$\hat{\omega} = \hat{h}(\tau,\zeta)\langle\zeta\rangle^{-s}\langle\tau - P(\zeta)\rangle^{b}.$$

Then

$$|\langle \Delta_x^{\frac{2b-1}{2}} \partial_x(uv), \omega\rangle| \leq I,$$

where I is defined by

$$I = \int_{\mathbb{R}^6} \frac{|\xi|\langle\xi\rangle^{2b-1}|\hat{f}(\tau-\tau',\zeta-\zeta')||\hat{g}(\tau',\zeta')||\hat{h}(\tau,\zeta)|\,d\tau d\zeta d\tau' d\zeta'}{\langle\sigma_2\rangle^{1/2}\langle\sigma_1\rangle^{1/2}\langle\sigma\rangle^{b}\langle\zeta-\zeta'\rangle^{s}\langle\zeta'\rangle^{s}\langle\zeta\rangle^{-s}}.$$

Recall (see [4]) that the following algebraic inequality holds:

$$3\max\left(|\sigma|, |\sigma_1|, |\sigma_2|\right) \geq |\xi\xi'(\xi'-\xi)| \geq |\xi|^2 \min\left(|\xi'|, |\xi'-\xi|\right). \tag{5.2}$$

Note also that

$$\frac{\langle\zeta\rangle^s}{\langle\zeta-\zeta'\rangle^s\langle\zeta'\rangle^s} \leq \frac{C}{\langle\zeta'\rangle^s} + \frac{C}{\langle\zeta-\zeta'\rangle^s}. \tag{5.3}$$

Now, to estimate I, by symmetry, we can always assume that $|\sigma_1| \geq |\sigma_2|$. Hence we have only to consider the two cases $|\sigma| \geq |\sigma_1|$ and $|\sigma_1| \geq |\sigma|$.

(a) The case $|\sigma| \geq |\sigma_1|$.

(a1) The subcase $|\xi'| \leq |\xi - \xi'|$. We denote by I_1 the contribution to I on this subdomain. From (5.2) we obtain

$$\frac{|\xi|\langle\xi\rangle^{2b-1}}{\langle\sigma\rangle^b} \leq \frac{|\xi|\langle\xi\rangle^{-1}}{|\xi'|^b} \leq |\xi'|^{-b},$$

and so from (5.3) we get

$$I_1 \leq \int_{\mathbb{R}^6} \frac{|\hat{f}(\tau-\tau',\zeta-\zeta')||\hat{g}(\tau',\zeta')||\hat{h}(\tau,\zeta)|}{\langle\sigma_2\rangle^{1/2}\langle\sigma_1\rangle^{1/2}|\xi'|^b\langle\zeta-\zeta'\rangle^s}\,d\tau d\zeta d\tau' d\zeta'$$

$$+\int_{\mathbb{R}^6}\frac{|\hat{f}(\tau-\tau',\zeta-\zeta')||\hat{g}(\tau',\zeta')||\hat{h}(\tau,\zeta)|}{\langle\sigma_2\rangle^{1/2}\langle\sigma_1\rangle^{1/2}|\xi'|^b\langle\zeta'\rangle^s}\,d\tau d\zeta d\tau' d\zeta'$$

$$\leq C\|\hat{f}(\tau,\zeta)\langle\zeta\rangle^{-s}\|_{L^2_{t,z}}\|\hat{g}\|_{L^2_{t,z}}\|\hat{h}\|_{L^2_{t,z}}+\|\hat{f}\|_{L^2_{t,z}}\|\hat{g}(\tau,\zeta)\langle\zeta\rangle^{-s}\|_{L^2_{t,z}}\|\hat{h}\|_{L^2_{t,z}}$$

by virtue of (3.11) in Lemma 3.4. Hence we have

$$I_1\leq\|u\|_{X^{1/2,0}}\|v\|_{X^{1/2,s}}\|\omega\|_{X^{b,-s}}+\|u\|_{X^{1/2,s}}\|v\|_{X^{1/2,0}}\|\omega\|_{X^{b,-s}}\ . \tag{5.4}$$

(a2) The subcase $|\xi-\xi'|\leq|\xi'|$. We denote by I_2 the contribution to I on this subdomain. From (5.2) again, we obtain

$$\frac{|\xi|\langle\xi\rangle^{2b-1}}{\langle\sigma\rangle^b}\leq\frac{|\xi|\langle\xi\rangle^{2b-1}}{\langle\xi\rangle^{2b}|\xi'-\xi|^{2b}}\leq|\xi'-\xi|^{-2b}.$$

Hence, from (5.3),

$$I_2\leq\int_{\mathbb{R}^6}\frac{|\hat{f}(\tau-\tau',\zeta-\zeta')||\hat{g}(\tau',\zeta')||\hat{h}(\tau,\zeta)|}{\langle\sigma_2\rangle^{1/2}\langle\sigma_1\rangle^{1/2}|\xi'-\xi|^b\langle\zeta-\zeta'\rangle^s}\,d\tau d\zeta d\tau' d\zeta'$$

$$+\int_{\mathbb{R}^6}\frac{|\hat{f}(\tau-\tau',\zeta-\zeta')||\hat{g}(\tau',\zeta')||\hat{h}(\tau,\zeta)|}{\langle\sigma_2\rangle^{1/2}\langle\sigma_1\rangle^{1/2}|\xi'-\xi|^b\langle\zeta'\rangle^s}\,d\tau d\zeta d\tau' d\zeta'$$

$$\leq C\|\hat{f}(\tau,\zeta)\langle\zeta\rangle^{-s}\|_{L^2_{t,z}}\|\hat{g}\|_{L^2_{t,z}}\|\hat{h}\|_{L^2_{t,z}}+\|\hat{f}\|_{L^2_{t,z}}\|\hat{g}(\tau,\zeta)\langle\zeta\rangle^{-s}\|_{L^2_{t,z}}\|\hat{h}\|_{L^2_{t,z}}$$

by virtue of (3.11) in Lemma 3.4. Hence we have

$$I_2\leq\|u\|_{X^{1/2,0}}\|v\|_{X^{1/2,s}}\|\omega\|_{X^{b,-s}}+\|u\|_{X^{1/2,s}}\|v\|_{X^{1/2,0}}\|\omega\|_{X^{b,-s}}. \tag{5.5}$$

(b) The case $|\sigma_1|\geq|\sigma|$. In this situation, from (5.2), we have

$$I=\int_{\mathbb{R}^6}\frac{|\xi|\langle\xi\rangle^{2b-1}|\hat{f}(\tau-\tau',\zeta-\zeta')||\hat{g}(\tau',\zeta')||\hat{h}(\tau,\zeta)|\,d\tau d\zeta d\tau' d\zeta'}{\langle\sigma_2\rangle^{1/2}\langle\sigma_1\rangle^{1/2-b}\langle\sigma_1\rangle^b\langle\sigma\rangle^b\langle\zeta-\zeta'\rangle^s\langle\zeta'\rangle^s\langle\zeta\rangle^{-s}}$$

$$\leq\int_{\mathbb{R}^6}\frac{|\hat{f}(\tau-\tau',\zeta-\zeta')||\hat{g}(\tau',\zeta')||\hat{h}(\tau,\zeta)|\langle\zeta\rangle^s\,d\tau d\zeta d\tau' d\zeta'}{\langle\sigma_2\rangle^{1/2}\langle\sigma_1\rangle^{1/2-b}\min\left(|\xi'|,|\xi-\xi'|\right)^b\langle\sigma\rangle^b\langle\zeta-\zeta'\rangle^s\langle\zeta'\rangle^s}.$$

(b1) Subcase $|\xi-\xi'|\leq|\xi'|$. We denote by I_3 the contribution to I on this subdomain. According to the last inequality and (5.3) and since $\langle\sigma\rangle\leq\langle\sigma_1\rangle$, we infer that

$$I_3\leq\int_{\mathbb{R}^6}\frac{|\hat{f}(\tau-\tau',\zeta-\zeta')||\langle\zeta'\rangle^{-s}\hat{g}(\tau',\zeta')||\hat{h}(\tau,\zeta)|}{\langle\sigma_2\rangle^{1/2}|\xi'-\xi|^b\langle\sigma\rangle^{1/2}}\,d\tau d\zeta d\tau' d\zeta'$$

$$+\int_{\mathbb{R}^6}\frac{|\langle\zeta-\zeta'\rangle^{-s}\hat{f}(\tau-\tau',\zeta-\zeta')||\hat{g}(\tau',\zeta')||\hat{h}(\tau,\zeta)|}{\langle\sigma_2\rangle^{1/2}|\xi'-\xi|^b\langle\sigma\rangle^{1/2}}\,d\tau d\zeta d\tau' d\zeta'$$

$$\leq C\|\hat{f}(\tau,\zeta)\langle\zeta\rangle^{-s}\|_{L^2_{t,z}}\|\hat{g}\|_{L^2_{t,z}}\|\hat{h}\|_{L^2_{t,z}}+\|\hat{f}\|_{L^2_{t,z}}\|\hat{g}(\tau,\zeta)\langle\zeta\rangle^{-s}\|_{L^2_{t,z}}\|\hat{h}\|_{L^2_{t,z}}$$

by virtue of (3.11) in Lemma 3.4. Hence we have

$$I_3\leq\|u\|_{X^{1/2,0}}\|v\|_{X^{1/2,s}}\|\omega\|_{X^{b,-s}}+\|u\|_{X^{1/2,s}}\|v\|_{X^{1/2,0}}\|\omega\|_{X^{b,-s}}. \tag{5.6}$$

(b2) Subcase $|\xi'| \le |\xi' - \xi|$. We denote by I_4 the contribution to I on this subdomain. Then according to (5.3) we have

$$I_4 \le \int_{\mathbb{R}^6} \frac{|\hat{f}(\tau-\tau',\zeta-\zeta')||\langle\zeta'\rangle^{-s}\hat{g}(\tau',\zeta')||\hat{\omega}(\tau,\zeta)|}{\langle\sigma_2\rangle^{1/2}|\xi'|^b\langle\sigma_1\rangle^{1/2-b}\langle\sigma\rangle^b}\, d\tau d\zeta d\tau' d\zeta'$$

$$+ \int_{\mathbb{R}^6} \frac{|\langle\zeta-\zeta'\rangle^{-s}\hat{f}(\tau-\tau',\zeta-\zeta')||\hat{g}(\tau',\zeta')||\hat{h}(\tau,\zeta)|}{\langle\sigma_2\rangle^{1/2}|\xi'|^b\langle\sigma_1\rangle^{1/2-b}\langle\sigma\rangle^b}\, d\tau d\zeta d\tau' d\zeta'$$

$$\le C\|\hat{f}(\tau,\zeta)\langle\zeta\rangle^{-s}\|_{L^2_{t,z}}\|\hat{g}\|_{L^2_{t,z}}\|\hat{h}\|_{L^2_{t,z}} + \|\hat{f}\|_{L^2_{t,z}}\|\hat{g}(\tau,\zeta)\langle\zeta\rangle^{-s}\|_{L^2_{t,z}}\|\hat{h}\|_{L^2_{t,z}}$$

by virtue of the estimate (3.12) in Lemma 3.4. Hence we have

$$I_4 \le \|u\|_{X^{1/2,0}}\|v\|_{X^{1/2,s}}\|\omega\|_{X^{b,-s}} + \|u\|_{X^{1/2,s}}\|v\|_{X^{1/2,0}}\|\omega\|_{X^{b,-s}}. \tag{5.7}$$

The proof follows now from (5.4), (5.5), (5.6), and (5.7). □

5.2. Local existence.

Proposition 5.2. *For any $\varphi \in H^s(\mathbb{R}^2)$ with $s \ge 0$, there exists $T = T(\|u_0\|_{L^2})$ and a unique local solution u of KPB* II *in*

$$Y_T = X_T^{1/2,s} \cap C([0,T], H^s).$$

Furthermore the map $\varphi \mapsto u$ is continuous from $H^s(\mathbb{R}^2)$ to Y_T.

Remark. It is important to note that the lower bound for the time of existence of u in $X_T^{1/2,s} \cap C([0,T], H^s)$ depends only on the norm of the corresponding initial data in L^2.

Proof. The procedure is the same as for the proof of Proposition 4.2. We now work in the space

$$Z = \{u \in X_T^{1/2,s} \;/\; \|u\|_Z = \|u\|_{X_T^{1/2,0}} + \gamma\,\|u\|_{X_T^{1/2,s}} < +\infty\},$$

where the constant γ is defined for all nontrivial φ by

$$\gamma = \frac{\|\varphi\|_{L^2}}{\|\varphi\|_{H^s}}.$$

From Propositions 2.3 and 2.5, for all $0 \le b < 1/2$,

$$\|L(u)\|_{X_T^{1/2,0}} \le C\|\varphi\|_{L^2} + C\|\Delta_x^{\frac{2b-1}{2}}\partial_x(u^2)\|_{X_T^{-b,0}},$$

$$\|L(u)\|_{X_T^{1/2,s}} \le C\|\varphi\|_{H^s} + C\|\Delta_x^{\frac{2b-1}{2}}\partial_x(u^2)\|_{X_T^{-b,s}}.$$

Choosing $b > 0$ small enough, Proposition 5.1 then yields

$$\|L(u)\|_{X_T^{1/2,0}} \le C\,\|\varphi\|_{L^2} + C\,T^\mu\|u\|^2_{X_T^{1/2,0}},$$

$$\|L(u)\|_{X_T^{1/2,s}} \le C\,\|\varphi\|_{H^s} + C\,T^\mu\|u\|^2_{X_T^{1/2,s}}.$$

Since $\partial_x(u^2) - \partial_x(v^2) = \partial_x[(u-v)(u+v)]$, in the same way we get

$$\|L(u) - L(v)\|_{X_T^{1/2,0}} \le C\,T^\mu \|u-v\|_{X_T^{1/2,0}} \|u+v\|_{X_T^{1/2,0}},$$

$$\|L(u) - L(v)\|_{X_T^{1/2,s}} \le C\,T^\mu \Big(\|u-v\|_{X_T^{1/2,s}} \|u+v\|_{X_T^{1/2,0}} + \|u-v\|_{X_T^{1/2,0}} \|u+v\|_{X_T^{1/2,s}} \Big).$$

Proceeding as in the proof of Proposition 4.2, it clearly follows from the above inequalities that for $T = (8C^2\|\varphi\|_{L^2})^{-\frac{1}{\mu}}$, the integral operator L is strictly contractive on the ball of radius $2C(\|\varphi\|_{L^2} + \gamma\|\varphi\|_{H^s})$ in Z. This mainly shows the result (we refer to the proof of Proposition 4.2 for further details). □

5.3. Global existence for KP II. We now prove the global existence result. In view of Proposition 5.2, for $\varphi \in H^s(\mathbb{R}^2)$, $s \ge 0$, the local solution u of KPB II can be extended on a maximal existence interval $[0, T_*[$ such that

$$\text{if } \ T_* < \infty\ , \text{ then } \ \limsup_{t \nearrow T_*} |u(t)|_2 = +\infty. \tag{5.8}$$

We are going to see that the L^2-norm of the solution is nonincreasing on $[0, T_*[$, which obviously ensures that $T_* = +\infty$.

Proof. Note that, thanks to Lemma 3.2 in [11], one can always find a sequence $(\varphi_n)_{n\ge 0} \subset \mathcal{N}$ such that $\varphi_n \to \varphi$ in $L^2(\mathbb{R}^2)$. Fixing $T \in]0, T_*[$, we deduce from Proposition 5.2 and Lemma 4.3 that for n large enough the solution u_n of KPB II with initial data φ_n satisfies (4.12) on $[0, T]$. We thus infer that

$$\|u_n(t_2)\|_2 \le \|u_n(t_1)\|_2\,, \quad 0 \le t_1 \le t_2 \le T.$$

Passing to the limit in n, using that $u_n \to u$ in $C([0,T], L^2(\mathbb{R}^2))$, we deduce that $t \mapsto \|u(t)\|_2$ is nonincreasing on $[0, T]$. Letting T tend to T_*, the result on $[0, T_*[$ is proved. □

Acknowledgments. We would like to thank N. Tzvetkov for useful discussions on Bourgain spaces, as well as the referees for valuable comments.

REFERENCES

[1] O. V. Besov, V. P. Il'in, and S. M. Nikolskii, *Integral Representations of Functions and Imbedding Theorems*, Vol. 1, J. Wiley, New York, 1978.

[2] J. Bourgain, *Fourier transform restriction phenomena for certain lattice subsets and application to nonlinear evolution equations* I. *Schrödinger equations*, Geom. Funct. Anal., 3 (1993), pp. 107–156.

[3] J. Bourgain, *Fourier transform restriction phenomena for certain lattice subsets and application to nonlinear evolution equations* II. *The KdV equation*, Geom. Funct. Anal., 3 (1993), pp. 209–262.

[4] J. Bourgain, *On the Cauchy problem for the Kadomtsev-Petviashvili equation*, Geom. Funct. Anal., 3 (1993), pp. 315–341.

[5] R. J. Iório, Jr., and W. V. L. Nunes, *On equations of KP-type*, Proc. Roy. Soc. Edinburgh A, 128 (1998), pp. 725–743.

[6] J. Ginibre, *Le problème de Cauchy pour des EDP semi-linéaires périodiques en variables d'espace (d'après Bourgain)*, Astérisque, 237 (1995), pp. 163–187.

[7] J. Ginibre and G. Velo, *Generalized Strichartz inequalities for the wave equation*, J. Funct. Anal., 151 (1997), pp. 384–436.

[8] J. Ginibre, Y. Tsutsumi, and G. Velo, *On the Cauchy problem for the Zakharov system*, J. Funct. Anal., 133 (1995), pp. 50–68.

[9] C. E. Kenig, G. Ponce, and L. Vega, *A bilinear estimate with applications to the KdV equation*, J. Amer. Math. Soc., 9 (1996), pp. 573–603.

[10] C. E. Kenig, G. Ponce, and L. Vega, *On the (generalized) Korteweg-de-Vries equation*, Duke Math. J., 59 (1989), pp. 585–610.

[11] L. Molinet, *On the asymptotic behavior of solutions to the (generalized) Kadomtsev-Petviashvili-Burgers equations*, J. Differential Equations, 152 (1999), pp. 30–74.

[12] L. Molinet, *The Cauchy problem for the (generalized) Kadomtsev-Petiashvili-Burgers equation*, Differential Integral Equations, 13 (2000), pp. 189–216.

[13] L. Molinet and F. Ribaud, *The Cauchy problem for dissipative Korteweg de Vries equations in Sobolev spaces of negative order*, Indiana Univ. Math. J., to appear.

[14] E. Ott and N. Sudan, *Damping of solitary waves*, Phys. Fluids, 13 (1970), pp. 1432–1434.

[15] J. C. Saut, *Remarks on the generalized Kadomtsev-Petviashvili equations*, Indiana Univ. Math. J., 42 (1993), pp. 1011–1026.

[16] H. Takaoka and N. Tzvetkov, *On the local regularity of Kadomtsev-Petviashvili*-II *equation*, Internat. Math. Res. Notices, 8 (2001), pp. 77–114.

SIAM J. MATH. ANAL.
Vol. 33, No. 6, pp. 1297–1304

STABILITY IN A LINEAR DELAY SYSTEM WITHOUT INSTANTANEOUS NEGATIVE FEEDBACK*

JOSEPH W.-H. SO[†], XIANHUA TANG[‡], AND XINGFU ZOU[§]

Abstract. It is shown that every solution of a linear differential system with constant coefficients and time delays tends to zero if a certain matrix derived from the coefficient matrix is a nonsingular M-matrix and the diagonal delays satisfy the so-called 3/2 condition.

Key words. linear system, pure-delay type, stability, diagonal dominant, M-matrix

AMS subject classifications. 34K20, 34K60

PII. S0036141001389263

1. Introduction.

Consider a system of delayed linear differential equations with constant coefficients of the form

$$\dot{x}_i(t) = -\sum_{j=1}^{n} a_{ij} x_j(t-\tau_{ij}), \qquad i = 1, 2, \dots, n, \tag{1.1}$$

with

$$\tau_{ij} \geq 0 \quad \text{for all} \quad 1 \leq i, j \leq n. \tag{1.2}$$

System (1.1) arises as linearization about an equilibrium point of many nonlinear systems with time delays. The interested reader can refer to Stépán [14] and the references therein for multiple-delay examples, such as machine tool vibration and human-machine systems.

When $\tau_{ij} = 0$ for all $i, j = 1, 2, \dots, n$, it is well known that (1.1) is asymptotically stable if and only if the matrix $A = (a_{ij})$ is a positively stable matrix, meaning that all eigenvalues of A have positive real parts. When some of the delays τ_{ij} are nonzero, (1.1) is asymptotically stable if and only if all the roots of its characteristic equation have negative real parts (cf. Hale and Verduyn Lunel [6]). In general, it is extremely difficult to analyze the characteristic equation of (1.1) when there are multiple (nonzero) delays. In Hofbauer and So [8], the authors considered the case when $\tau_{ii} = 0$ for $i = 1, 2, \dots, n$, and they established the following result.

THEOREM 1.1. *Assume that $\tau_{ii} = 0$ for all $i = 1, 2, \dots, n$. Then* (1.1) *is asymptotically stable for all choices of delays of the form* (1.2) *if and only if $a_{ii} > 0$ for $i = 1, 2, \dots, n$, $\det A \neq 0$, and A is weakly diagonally dominant (i.e., all the principal minors of $\hat{A} = (\hat{a}_{ij})$ are nonnegative, where $\hat{a}_{ii} = a_{ii}$ and $\hat{a}_{ij} = -|a_{ij}|$ for $j \neq i$).*

In such a case (i.e., when there is no diagonal delay), Györi [5] also obtained a similar result for a quasi-monotone matrix A (i.e., $a_{ij} \leq 0$ for $i \neq j$). Motivated by the

[*] Received by the editors May 10, 2001; accepted for publication (in revised form) October 16, 2001; published electronically April 26, 2002. This work was supported by NNSF of China and NSERC of Canada.

http://www.siam.org/journals/sima/33-6/38926.html

[†] Department of Mathematical Sciences, University of Alberta, Edmonton, AB, Canada T6G 2G1 (joseph.so@ualberta.ca).

[‡] Department of Mathematics, Central South University, Changsha, Hunan 410083, People's Republic of China (xhtang@public.cs.hn.cn).

[§] Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NF, Canada A1C 5S7 (xzou@math.mun.ca).

study of neural networks of Hopfield type, there is a recent extension of Theorem 1.1 by Sue Ann Campbell [2] of the University of Waterloo to include both types of diagonal terms, both with and without delays. A result similar to Theorem 1.1 was obtained (with the conditions on a suitably derived matrix) using the same proof as in [8].

When $\tau_{ii} \neq 0$, $i = 1, \dots, n$, instantaneous feedback is absent, and (1.1) becomes a system of "pure-delay type." For such a "pure-delay-type" system, the stability problem becomes much harder, as pointed out by Gopalsamy and He [4], He [7], and Kuang [10]. However, it is reasonable to expect that a similar stability criterion holds as long as the diagonal delays are sufficiently small. This paper will provide an answer to this question. More precisely, by employing a new technique (without analyzing the characteristic equation or constructing a Liapunov functional), we will extend the sufficiency part of Theorem 1.1 to the case when τ_{ii} $(i = 1, 2, \dots, n)$ are not necessarily all zero. For convenience, we recall the concept of a nonsingular M-matrix (cf. Fiedler [3]).

Definition 1.2. *The $n \times n$ matrix $B = (b_{ij})$ is a nonsingular M-matrix if* (i) $b_{ij} \leq 0$ *for* $j \neq i$ *and* (ii) *all principal minors of B are positive.*

There are many equivalent formulations of this concept (cf. Fiedler [3, Theorem 5.1, p. 114]. In particular, if B is a nonsingular M-matrix, then B^{-1} is a positive matrix.

We associate with the $n \times n$ matrix $A = (a_{ij})$ a new matrix $\tilde{A} = (\tilde{a}_{ij})$ defined by

$$\tilde{a}_{ii} = a_{ii} \quad \text{for} \quad i = 1, 2, \dots, n \tag{1.3}$$

and

$$\tilde{a}_{ij} = -\frac{1 + \frac{1}{9} a_{ii}\tau_{ii}(3 + 2a_{ii}\tau_{ii})}{1 - \frac{1}{9} a_{ii}\tau_{ii}(3 + 2a_{ii}\tau_{ii})} |a_{ij}| \quad \text{for} \quad i \neq j, \quad j = 1, 2, \dots, n. \tag{1.4}$$

Now we can state our main result.

Theorem 1.3. *Assume that*

$$a_{ii}\tau_{ii} < \frac{3}{2} \quad \textit{for all} \quad i = 1, 2, \dots, n. \tag{1.5}$$

If $\tilde{A}$ is a nonsingular M-matrix, then every solution $(x_1(t), x_2(t), \dots, x_n(t))$ of (1.1) *tends to* 0 *as* $t \to \infty$.

Remark 1.1. Condition (1.5) will be referred to as the 3/2 condition. When $\tau_{ii} = 0$ for all $i = 1, \dots, n$, the 3/2 condition is automatically satisfied and $\tilde{A} = \hat{A}$. According to Bapat and Raghavan [1, Theorem 7.8.6], if $\hat{A}$ is a nonsingular M-matrix, then A itself is nonsingular. Hence, in the case of no diagonal delays, a matrix A satisfying the hypotheses of Theorem 1.3 will also satisfy the criterion in Theorem 1.1. The stability criterion in Theorem 1.3 is concrete and easily verifiable for any given (numerical) system.

Remark 1.2. There are many 3/2 stability results for scalar (linear or nonlinear, autonomous or nonautonomous, one or several delays) equations in the literature. See, for example, [11, 16, 15, 9, 12, 13]. It would be interesting to see if these results can be extended to systems.

Remark 1.3. In [8], besides the linear equation (1.1), the authors also considered global stability of Lotka–Volterra equations (with $\tau_{ii} = 0$). We are currently investigating the possibility of a 3/2 result for Lotka–Volterra systems when $\tau_{ii} > 0$.

2. Proof of Theorem 1.3. The proof of Theorem 1.3 consists of the following two lemmas. The first lemma establishes the boundedness of solutions of (1.1).

LEMMA 2.1. *Under the conditions of Theorem* 1.3, *every (forward) solution of* (1.1) *is bounded.*

Proof. Let $(x_1(t), x_2(t), \ldots, x_n(t))$ be a solution of (1.1) on $[t_0, \infty)$. Without loss of generality, t_0 can be taken to be 0. For the sake of contradiction, assume that $\max\{|x_i(t)| : i = 1, 2, \ldots, n\}$ is unbounded on $[t_0, \infty)$. By rearranging the indices i, we may assume that

$$\limsup_{t\to\infty} |x_i(t)| = \infty \quad \text{for} \quad i = 1, 2, \ldots, k(\leq n) \tag{2.1}$$

and

$$|x_i(t)| \leq M \quad \text{for} \quad t \geq t_0 - \max_{h,k}\{\tau_{hk}\}, \qquad i = k+1, \ldots, n. \tag{2.2}$$

Let N be the smallest integer such that $N > t_0 + \tau_{ii}$ for all i. There is an integer $N_1 > N$ such that for each $i = 1, \ldots, k$, the maximum of the function $|x_i(t)|$ on the interval $[t_0, N_1]$ is attained at a point in $[N, N_1]$. Fix $i = 1, \ldots, k$. For each integer $m \geq 1$, let $t_{im} \in [N, N_1 + m]$ be such that $|x_i(t_{im})| = \max\{|x_i(t)| : t \in [t_0, N_1 + m]\}$. We may assume that $\{t_{im}\}_{m=1}^{\infty}$ is a nondecreasing sequence. By going to subsequences if necessary, we have k sequences $\{t_{im}\}_{m=1}^{\infty}, i = 1, 2, \ldots, k$, such that

$$\left\{ \begin{array}{l} t_{im} \uparrow \infty, \quad |x_i(t_{im})| \uparrow \infty \quad \text{as} \quad m \to \infty, \\ |x_i(t)| \leq |x_i(t_{im})| \quad \text{for} \quad t_0 \leq t \leq t_m, \end{array} \right. \qquad \text{for} \quad i = 1, 2, \ldots, k, \tag{2.3}$$

where $t_m = \max\{t_{im} : i = 1, 2, \ldots, k\}$. Again by going to subsequences if necessary, we may assume that for each $i = 1, \ldots, k$, all the terms in the sequence $\{x_i(t_{im})\}_{m=1}^{\infty}$ are of the same sign. Without loss of generality (i.e., by using $-x_i(t)$ instead of $x_i(t)$ and $-a_{ij}$ instead of a_{ij} for $j \neq i$, if necessary), we may assume that $|x_i(t_{im})| = x_i(t_{im})$. Then

$$|x_i(t)| \leq x_i(t_{im}) \quad \text{for} \quad t_0 \leq t < t_m \qquad \text{and} \qquad \dot{x}_i(t_{im}) \geq 0, \qquad i = 1, 2, \ldots, k.$$

It follows from (1.1) that

$$0 \leq -\sum_{j=1}^{n} a_{ij} x_j(t_{im} - \tau_{ij}) \leq -a_{ii} x_i(t_{im} - \tau_{ii}) + \sum_{j\neq i}^{k} |a_{ij}| x_j(t_{jm}) + M \sum_{j=k+1}^{n} |a_{ij}|$$

or

$$x_i(t_{im} - \tau_{ii}) \leq \frac{1}{a_{ii}} \left[\sum_{j\neq i}^{k} |a_{ij}| x_j(t_{jm}) + M \sum_{j=k+1}^{n} |a_{ij}| \right], \qquad i = 1, 2, \ldots, k. \tag{2.4}$$

Set

$$\alpha_i = \frac{1}{a_{ii}} \left[\sum_{j\neq i}^{k} |a_{ij}| x_j(t_{jm}) + M \sum_{j=k+1}^{n} |a_{ij}| \right], \qquad i = 1, 2, \ldots, k. \tag{2.5}$$

We will now show

$$a_{ii} x_i(t_{im}) + \sum_{j\neq i}^{k} \tilde{a}_{ij} x_j(t_{jm}) \leq M \sum_{j=k+1}^{n} |\tilde{a}_{ij}| \quad \text{for} \quad i = 1, 2, \ldots, k. \tag{2.6}$$

If $x_i(t_{im}) \leq \alpha_i$, then (2.6) follows from a simple calculation. If $x_i(t_{im}) > \alpha_i$, by (2.4) there exists $\xi_{im} \in [t_{im} - \tau_{ii}, t_{im}]$ such that $x_i(\xi_{im}) = \alpha_i$. From (1.1) we have

$$\dot{x}_i(t) \leq a_{ii}[-x_i(t-\tau_{ii}) + \alpha_i] \leq a_{ii}\left(|x_i(t_{im})| + \alpha_i\right) \quad \text{for} \quad N \leq t \leq t_m. \tag{2.7}$$

For $t \in [\xi_{im}, t_{im})$, integrating (2.7) from $t - \tau_{ii}$ to ξ_{im}, we have

$$\alpha_i - x_i(t-\tau_{ii}) \leq a_{ii}\left(|x_i(t_{im})| + \alpha_i\right)(\xi_{im} + \tau_{ii} - t) \quad \text{for} \quad \xi_{im} \leq t \leq t_{im}.$$

Substituting this into the first inequality in (2.7), we obtain

$$\dot{x}_i(t) \leq a_{ii}^2\left(|x_i(t_{im})| + \alpha_i\right)(\xi_{im} + \tau_{ii} - t) \quad \text{for} \quad \xi_{im} \leq t \leq t_{im}.$$

Combining this and (2.7), we have

$$\dot{x}_i(t) \leq a_{ii}\left(|x_i(t_{im})| + \alpha_i\right)\min\{1, a_{ii}(\xi_{im} + \tau_{ii} - t)\} \quad \text{for} \quad \xi_{im} \leq t \leq t_{im}. \tag{2.8}$$

We consider the following two cases.

Case 1. $t_{im} - \xi_{im} \leq 2\tau_{ii}/3$. In this case, by (2.8) we have

$$\begin{aligned} x_i(t_{im}) - x_i(\xi_{im}) &\leq a_{ii}^2\left(|x_i(t_{im})| + \alpha_i\right)\int_{\xi_{im}}^{t_{im}} (\xi_{im} + \tau_{ii} - t)dt \\ &= a_{ii}^2\left(|x_i(t_{im})| + \alpha_i\right)\left[\tau_{ii}(t_{im} - \xi_{im}) - \frac{1}{2}(t_{im} - \xi_{im})^2\right] \\ &\leq \left(|x_i(t_{im})| + \alpha_i\right)\left[\frac{2}{3}(a_{ii}\tau_{ii})^2 - \frac{2}{9}(a_{ii}\tau_{ii})^2\right] \\ &= \frac{4}{9}(a_{ii}\tau_{ii})^2\left(|x_i(t_{im})| + \alpha_i\right) \\ &\leq \frac{1}{9}a_{ii}\tau_{ii}(3 + 2a_{ii}\tau_{ii})\left(|x_i(t_{im})| + \alpha_i\right), \end{aligned}$$

since the function $y \mapsto \tau_{ii}y - \frac{1}{2}y^2$ is increasing on the interval $[0, \frac{2\tau_{ii}}{3}]$.

Case 2. $t_{im} - \xi_{im} > 2\tau_{ii}/3$. In this case, let $t_{im} - \eta_{im} = 2\tau_{ii}/3$ so that $\eta_{im} \in (\xi_{im}, t_{im}]$. Then by (2.8) we have

$$\begin{aligned} &x_i(t_{im}) - x_i(\xi_{im}) \\ &\leq \left(|x_i(t_{im})| + \alpha_i\right)\left[a_{ii}(\eta_{im} - \xi_{im}) + a_{ii}^2\int_{\eta_{im}}^{t_{im}} (\xi_{im} + \tau_{ii} - t)dt\right] \\ &= \left(|x_i(t_{im})| + \alpha_i\right)\left[a_{ii}(\eta_{im} - \xi_{im})(1 - a_{ii}(t_{im} - \eta_{im})) + a_{ii}^2\int_{\eta_{im}}^{t_{im}} (\eta_{im} + \tau_{ii} - t)dt\right] \\ &= \left(|x_i(t_{im})| + \alpha_i\right)\left[a_{ii}(\eta_{im} - \xi_{im})\left(1 - \frac{2}{3}a_{ii}\tau_{ii}\right)\right. \\ &\qquad\qquad \left. + \, a_{ii}^2\tau_{ii}(t_{im} - \eta_{im}) - \frac{1}{2}a_{ii}^2(t_{im} - \eta_{im})^2\right] \\ &\leq \left(|x_i(t_{im})| + \alpha_i\right)\left[\frac{1}{3}a_{ii}\tau_{ii} + \frac{2}{9}(a_{ii}\tau_{ii})^2\right] \\ &= \frac{1}{9}a_{ii}\tau_{ii}(3 + 2a_{ii}\tau_{ii})\left(|x_i(t_{im})| + \alpha_i\right), \end{aligned}$$

since $\eta_{im} - \xi_{im} \leq \frac{\tau_{ii}}{3}$ in this case.

Combining Cases 1 and 2, we have

$$\begin{aligned} &a_{ii}x_i(t_{im}) \\ &\le \frac{1+\frac{1}{9}a_{ii}\tau_{ii}(3+2a_{ii}\tau_{ii})}{1-\frac{1}{9}a_{ii}\tau_{ii}(3+2a_{ii}\tau_{ii})}\left[\sum_{j\neq i}^{k}|a_{ij}|x_j(t_{jm})+M\sum_{j=k+1}^{n}|a_{ij}|\right], \qquad i=1,2,\dots,k, \end{aligned}$$

which implies (2.6) is true.

Let $\tilde{A}_k=(\tilde{a}_{ij})_{k\times k}$ denote the kth leading principal submatrix of $\tilde{A}$. Then $\tilde{A}_k$ is a nonsingular M-matrix of order k, and so $\tilde{A}_k^{-1}>0$. Hence, it follows from (2.6) that

$$(x_1(t_{1m}),x_2(t_{2m}),\dots,x_k(t_{km}))^T\le M\tilde{A}_k^{-1}\left(\sum_{j=k+1}^{n}|\tilde{a}_{1j}|,\sum_{j=k+1}^{n}|\tilde{a}_{2j}|,\dots,\sum_{j=k+1}^{n}|\tilde{a}_{kj}|\right)^T, \quad m=1,2,\dots.$$

We conclude that

$$\limsup_{m\to\infty}|x_i(t_{im})|<\infty, \qquad i=1,2,\dots,k.$$

This contradicts the fact that $|x_i(t_{im})|\to\infty$ as $m\to\infty$ for $i=1,2,\dots,k$, and the proof is complete.

Next, using the boundedness of solutions, we can prove the convergence of all solutions of (1.1).

LEMMA 2.2. *Under the conditions of Theorem* 1.3*, every solution of* (1.1) *tends* 0 *as* $t\to\infty$.

Proof. Let $(x_1(t),x_2(t),\dots,x_n(t))$ be a solution of (1.1) on $[t_0,\infty)$. We will prove that

$$\lim_{t\to\infty}x_i(t)=0, \qquad i=1,2,\dots,n. \tag{2.9}$$

We distinguish the two cases.

Case A. All of the functions $\sum_{j=1}^{n}a_{ij}x_j(t-\tau_{ij})$, $i=1,2,\dots,n$, are nonoscillatory. Then the functions $\dot{x}_i(t)$ $(i=1,2,\dots,n)$ are eventually sign-definite, and so by Lemma 2.1, the limit $c_i=\lim_{t\to\infty}x_i(t)$ exists. By (1.1), $\dot{x}_i(t)$ converges as $t\to\infty$. Since $\dot{x}_i(t)$ is bounded, $x_i(t)$ is uniformly continuous and convergent. Therefore, $\lim_{t\to\infty}\dot{x}_i(t)=0$ for $i=1,2,\dots,n$, and we have

$$\sum_{j=1}^{n}a_{ij}c_j=0 \quad\text{for}\quad i=1,2,\dots,n.$$

It follows that

$$a_{ii}|c_i|-\sum_{j\neq i}|a_{ij}||c_j|\le 0 \quad\text{for}\quad i=1,2,\dots,n. \tag{2.10}$$

Set $\hat{A}=(\hat{a}_{ij})$, where $\hat{a}_{ii}=a_{ii}$ and $\hat{a}_{ij}=-|a_{ij}|$ for $j\neq i$. Then $\hat{A}\ge\tilde{A}$ and $\hat{A}$ has nonpositive off-diagonal entries. In view of [3, Theorem 2.5.4], the matrix $\hat{A}$ is also a nonsingular M-matrix. Since (2.10) can be expressed as the matrix inequality $\hat{A}(|c_1|,\dots,|c_n|)^T\le(0,\dots,0)^T$, by applying the positive matrix $\hat{A}^{-1}$ to both sides, we conclude that $c_1=c_2=\dots=c_n=0$.

Case B. At least one of the functions $\sum_{j=1}^{n} a_{ij}x_j(t-\tau_{ij})$ $(i = 1, 2, \ldots, n)$ is oscillatory. Set

$$U_i = \limsup_{t\to\infty} |x_i(t)|, \qquad i = 1, 2, \ldots, n.$$

By Lemma 2.1, we have $U_i < \infty$, $i = 1, 2, \ldots, n$. It suffices to prove that $U_1 = \cdots = U_n = 0$. By rearranging the indices, we may assume that $\sum_{j=1}^{n} a_{ij}x_j(t-\tau_{ij})$, $i = 1, \ldots, k$, are oscillatory and $\sum_{j=1}^{n} a_{ij}x_j(t-\tau_{ij})$, $i = k+1, \ldots, n$, are nonoscillatory. It follows from (1.1) that $\dot{x}_i(t)$ $(i = 1, 2, \ldots, k)$ are oscillatory and

$$\lim_{t\to\infty} \dot{x}_i(t) = 0 \quad \text{for} \quad i = k+1, \ldots, n. \tag{2.11}$$

Hence, for any $\epsilon > 0$, there exist k sequences $\{t_{im}\}$ $i = 1, 2, \ldots, k$, such that

$$\begin{cases} t_{im} \uparrow \infty, \quad |x_i(t_{im})| \to U_i \quad \text{as} \quad m \to \infty, \\ |\dot{x}_i(t_{im})| = 0, \quad U_i - \epsilon < |x_i(t)| < U_i + \epsilon \quad \text{for} \quad t \ge t_1, \end{cases} \qquad i = 1, 2, \ldots, k, \tag{2.12}$$

where $t_1 = \min\{t_{i1} : i = 1, 2, \ldots, k\}$. By going to subsequences if necessary, we may assume $|x_i(t_{im})| = x_i(t_{im})$ (use $-x_i(t)$ instead of $x_i(t)$ and $-a_{ij}$ instead of a_{ij} for $j \ne i$, if necessary). By (1.1), as long as m is sufficiently large, we have

$$0 = -\sum_{j=1}^{n} a_{ij}x_j(t_{im}-\tau_{ij}) \le -a_{ii}x_i(t_{im}-\tau_{ii}) + \sum_{j\ne i}^{n} |a_{ij}|(U_j+\epsilon)$$

or

$$x_i(t_{im}-\tau_{ii}) \le \frac{1}{a_{ii}} \sum_{j\ne i}^{n} |a_{ij}|(U_j+\epsilon), \qquad i = 1, 2, \ldots, k. \tag{2.13}$$

Set

$$\beta_i = \frac{1}{a_{ii}} \sum_{j\ne i}^{n} |a_{ij}|(U_j+\epsilon), \qquad i = 1, 2, \ldots, k. \tag{2.14}$$

We will now show

$$a_{ii}x_i(t_{im}) + \sum_{j\ne i} \tilde{a}_{ij}(U_j+\epsilon) \le \frac{2\epsilon a_{ii}\tau_{ii}(3+2a_{ii}\tau_{ii})}{9-a_{ii}\tau_{ii}(3+2a_{ii}\tau_{ii})}, \qquad i = 1, 2, \ldots, k. \tag{2.15}$$

If $x_i(t_{im}) \le \beta_i$, then (2.15) obviously holds. If $x_i(t_{im}) > \beta_i$, by (2.13) there exists $\xi_{im} \in [t_{im}-\tau_{ii}, t_{im}]$ such that $x_i(\xi_{im}) = \beta_i$. Using (1.1), for m sufficiently large we have

$$\dot{x}_i(t) \le a_{ii}[-x_i(t-\tau_{ii}) + \beta_i] \le a_{ii}\left[(U_i+\epsilon) + \beta_i\right] \quad \text{for} \quad \xi_{im} - \tau_{ii} \le t \le t_{im}. \tag{2.16}$$

For $t \in [\xi_{im}, t_{im})$, integrating (2.16) from $t-\tau_{ii}$ to ξ_{im}, we have

$$\beta_i - x_i(t-\tau_{ii}) \le a_{ii}\left[(U_i+\epsilon)+\beta_i\right](\xi_{im}+\tau_{ii}-t) \quad \text{for} \quad \xi_{im} \le t \le t_{im}.$$

Substituting this into the first inequality in (2.16), we obtain

$$\dot{x}_i(t) \le a_{ii}^2\left[(U_i+\epsilon)+\beta_i\right](\xi_{im}+\tau_{ii}-t) \quad \text{for} \quad \xi_{im} \le t \le t_{im}.$$

Combining this and (2.16), we have

$$\dot{x}_i(t) \le a_{ii}\left[(U_i+\epsilon)+\beta_i\right]\min\{1, a_{ii}(\xi_{im}+\tau_{ii}-t)\}, \qquad \xi_{im}\le t\le t_{im}. \tag{2.17}$$

We consider the following two cases.

Case 1. $t_{im}-\xi_{im}\le 2\tau_{ii}/3$. In this case, by (2.17) we have

$$\begin{aligned}
x_i(t_{im})-x_i(\xi_{im}) &\le a_{ii}^2\left[(U_i+\epsilon)+\beta_i\right]\int_{\xi_{im}}^{t_{im}}(\xi_{im}+\tau_{ii}-t)dt\\
&= a_{ii}^2\left[(U_i+\epsilon)+\beta_i\right]\left[\tau_{ii}(t_{im}-\xi_{im})-\frac{1}{2}(t_{im}-\xi_{im})^2\right]\\
&\le \left[(U_i+\epsilon)+\beta_i\right]\left[\frac{2}{3}(a_{ii}\tau_{ii})^2-\frac{2}{9}(a_{ii}\tau_{ii})^2\right]\\
&= \frac{4}{9}(a_{ii}\tau_{ii})^2\left[(U_i+\epsilon)+\beta_i\right]\\
&\le \frac{1}{9}a_{ii}\tau_{ii}(3+2a_{ii}\tau_{ii})\left[(U_i+\epsilon)+\beta_i\right]\\
&= \frac{1}{9}a_{ii}\tau_{ii}(3+2a_{ii}\tau_{ii})\left[(U_i-\epsilon)+\beta_i+2\epsilon\right]
\end{aligned}$$

by the 3/2 condition (1.5).

Case 2. $t_{im}-\xi_{im}>2\tau_{ii}/3$. In this case, let $t_{im}-\eta_{im}=2\tau_{ii}/3$. Then $\eta_{im}\in(\xi_{im},t_{im}]$. By (2.17), we have

$$\begin{aligned}
&x_i(t_{im})-x_i(\xi_{im})\\
&\le \left[(U_i+\epsilon)+\beta_i\right]\left[a_{ii}(\eta_{im}-\xi_{im})+a_{ii}^2\int_{\eta_{im}}^{t_{im}}(\xi_{im}+\tau_{ii}-t)dt\right]\\
&= \left[(U_i+\epsilon)+\beta_i\right]\left[a_{ii}(\eta_{im}-\xi_{im})(1-a_{ii}(t_{im}-\eta_{im}))+a_{ii}^2\int_{\eta_{im}}^{t_{im}}(\eta_{im}+\tau_{ii}-t)dt\right]\\
&= \left[(U_i+\epsilon)+\beta_i\right]\left[a_{ii}(\eta_{im}-\xi_{im})\left(1-\frac{2}{3}a_{ii}\tau_{ii}\right)+a_{ii}^2\tau_{ii}(t_{im}-\eta_{im})-\frac{1}{2}a_{ii}^2(t_{im}-\eta_{im})^2\right]\\
&\le \left[(U_i+\epsilon)+\beta_i\right]\left[\frac{1}{3}a_{ii}\tau_{ii}+\frac{2}{9}(a_{ii}\tau_{ii})^2\right]\\
&= \frac{1}{9}a_{ii}\tau_{ii}(3+2a_{ii}\tau_{ii})\left[(U_i+\epsilon)+\beta_i\right]\\
&= \frac{1}{9}a_{ii}\tau_{ii}(3+2a_{ii}\tau_{ii})\left[(U_i-\epsilon)+\beta_i+2\epsilon\right],
\end{aligned}$$

since $\eta_{im}-\xi_{im}\le\frac{\tau_{ii}}{3}$.

Combining Cases 1 and 2 with (2.12), we have

$$\begin{aligned}
&a_{ii}x_i(t_{im})\\
&\le \frac{1+\frac{1}{9}a_{ii}\tau_{ii}(3+2a_{ii}\tau_{ii})}{1-\frac{1}{9}a_{ii}\tau_{ii}(3+2a_{ii}\tau_{ii})}\sum_{j\ne i}|a_{ij}|(U_i+\epsilon)+\frac{2\epsilon a_{ii}\tau_{ii}(3+2a_{ii}\tau_{ii})}{9-a_{ii}\tau_{ii}(3+2a_{ii}\tau_{ii})}, \qquad i=1,2,\ldots,k.
\end{aligned}$$

This shows (2.15) is true. Letting $m\to\infty$ and $\epsilon\to 0$ in (2.15), we obtain

$$a_{ii}U_i+\sum_{j\ne i}\tilde{a}_{ij}U_j\le 0 \quad \text{for} \quad i=1,2,\ldots,k. \tag{2.18}$$

On the other hand, for each $i = k+1, \dots, n$, let $\{s_{im}\}_{m=1}^{\infty} \uparrow \infty$ be such that $\lim_{m\to\infty} x_i(s_{im}) = U_i$. By (2.11), we have $\lim_{m\to\infty} \dot{x}_i(s_{im} + \tau_{ii}) = 0$. Using (1.1), we have

$$\begin{aligned} 0 &= \dot{x}_i(s_{im} + \tau_{ii}) + a_{ii}x_i(s_{im}) + \sum_{j\neq i} a_{ij}x_j(s_{im} + \tau_{ii} - \tau_{ij}) \\ &\geq \dot{x}_i(s_{im} + \tau_{ii}) + a_{ii}x_i(s_{im}) + \sum_{j\neq i} \tilde{a}_{ij}|x_j(s_{im} + \tau_{ii} - \tau_{ij})|, \end{aligned}$$

since $\tilde{a}_{ij} \leq -|a_{ij}| \leq 0$. Letting $m \to \infty$, we obtain

$$a_{ii}U_i + \sum_{j\neq i} \tilde{a}_{ij}U_j \leq 0 \quad \text{for} \quad i = k+1, \dots, n. \tag{2.19}$$

By (2.17) and (2.18) and using the fact that $\tilde{A}$ is a nonsingular M-matrix (so that $\tilde{A}^{-1}$ is a positive matrix), we have $U_1 = U_2 = \cdots = U_n = 0$. The proof is now complete.

Acknowledgments. The authors would like to thank the two anonymous referees for their useful suggestions regarding additional references.

REFERENCES

[1] R. B. Bapat and T. E. S. Raghavan, *Nonnegative Matrices and Applications*, Cambridge University Press, Cambridge, UK, 1997.

[2] S. A. Campbell, *private communication*, University of Waterloo, Waterloo, ON, Canada.

[3] M. Fiedler, *Special Matrices and Their Applications in Numerical Mathematics*, Martinus Nijhoff, Dordrecht, The Netherlands, 1986.

[4] K. Gopalsamy and X. He, *Global stability in n-species competition modelled by "pure-delay type" systems* II*: Nonautonomous case*, Canad. Appl. Math. Quart., 6 (1998), pp. 17–43.

[5] I. Györi, *Stability in a class of integrodifferential systems*, in Recent Trends in Differential Equations, World Sci. Ser. Appl. Anal. 1, R. P. Agarwal, ed., World Scientific, Singapore, 1992, pp. 269–284.

[6] J. K. Hale and S. M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.

[7] X. He, *Global stability in nonautonomous Lotka-Volterra systems of "pure-delay type,"* Differential Integral Equations, 11 (1998), pp. 293–310.

[8] J. Hofbauer and J. W.–H. So, *Diagonal dominance and harmless off-diagonal delays*, Proc. Amer. Math. Soc., 128 (2000), pp. 2675–2682.

[9] T. Krisztin, *On stability properties for one-dimensional functional differential equations*, Funkcial. Ekvac., 34 (1991), pp. 241–256.

[10] Y. Kuang, *Global stability in delay differential systems without dominating instantaneous negative feedbacks*, J. Differential Equations, 119 (1995), pp. 503–532.

[11] A. D. Myshkis, *Linear Differential Equations with Retarded Arguments*, Nauka, Moscow, 1972 (in Russian).

[12] J. W.-H. So and J. S. Yu, *Global attractivity for a population model with time delay*, Proc. Amer. Math. Soc., 123 (1995), pp. 2687–2694.

[13] J. W.-H. So and J. S. Yu, *Global stability of a general population model with time delays*, Fields Inst. Commun., 21 (1999), pp. 447–457.

[14] G. Stépán, *Retarded Dynamical Systems: Stability and Characteristic Functions*, Pitman Res. Notes Math. Ser. 210, Longman Scientific & Technical, Harlow, UK, 1989.

[15] T. Yoneyama, *On the $\frac{3}{2}$ stability theorem for one-dimensional delay-differential equations*, J. Math. Anal. Appl., 125 (1987), pp. 161–173.

[16] J. A. Yorke, *Asymptotic stability for one dimensional differential-delay equations*, J. Differential Equations, 7 (1970), pp. 189–202.

SIAM J. MATH. ANAL.
Vol. 33, No. 6, pp. 1305–1319

PATH-TRACKING THROUGH SINGULARITIES*

G. W. REDDIEN†

Abstract. Given a mapping F from R^n to R^m with $n \geq m$, a basic computational problem that has applications in robotics is to solve for paths $x(t)$ in the domain of F given a target path $z(t)$ in the range of F so that $F(x(t)) = z(t)$. Results are given for the existence and characterization of solution paths through singularities of F, with emphasis on folds and cusps. Reparametrizations using Puiseux series are derived in cases where $x(t)$ is not differentiable.

Key words. singularities, Puiseux series, robotics

AMS subject classifications. Primary, 65H10; Secondary, 41A58

PII. S0036141000373872

1. Introduction.

A motivating example for the problems treated here is a three-bar mechanism, or arm, constrained to move in the plane. Given joint lengths l_1, l_2, l_3 and joint angles θ_1, θ_2, and θ_3 as indicated in Figure 1, the coordinates of the point P are determined. Indeed, it follows that coordinates are given by the formulas

$$\begin{aligned} z_1 &= l_1 \cos\theta_1 + l_2 \cos(\theta_1 + \theta_2) + l_3 \cos(\theta_1 + \theta_2 + \theta_3), \\ z_2 &= l_1 \sin\theta_1 + l_2 \sin(\theta_1 + \theta_2) + l_3 \sin(\theta_1 + \theta_2 + \theta_3) \end{aligned} \tag{1.1}$$

defining a mapping F from R^3 to R^2. If the motion of the links is constrained by $a_i \leq \theta_i \leq b_i$, $i = 1, 2, 3$, then the change of variables

$$\theta_i = \frac{a_i + b_i}{2} + \frac{1}{2}(b_i - a_i)\sin x_i, \qquad i = 1, 2, 3, \tag{1.2}$$

produces $F = F(x)$, where x lies in any open set D containing $[-\pi/2, \pi/2]^3$. Although simple in form, this device contains many of the kinematic singularities important in robotics. An example of a spatial manipulator is given in section 4.

Let $F : R^n \to R^m$ with $n \geq m$ be a smooth mapping defined on an open, connected set $D \subset R^n$, and let $z_0 = F(x_0)$. We call z_0 a boundary point (local) for $F(D)$ if there exists an open neighborhood U of x_0 so that z_0 is not in the interior of $F(U)$. Equivalently, we could say that z_0 is a boundary point of F at $F(x_0) = z_0$ if F is not open at x_0. The system of equations

$$\hat{F}(x, z) = F(x) - z = 0 \tag{1.3}$$

represents m-equations in $(n + m)$-unknowns. If $\hat{F}_x(x_0, z_0) = F_x(x_0)$ has full rank, then the implicit function theorem guarantees that (1.3) can be solved for all z near z_0; that is, if $F_x(x_0)$ has full rank, then $z_0 = F(x_0)$ is not a boundary point of $F(D)$. Thus the problem of determining the boundary of $F(D)$ is closely related to the problem of finding points where F_x drops in rank. We assume throughout this paper that

$$\operatorname{rank}(F_x) \geq m - 1 \text{ on } D, \tag{H1}$$

[*] Received by the editors June 14, 2000; accepted for publication (in revised form) September 11, 2001; published electronically April 26, 2002.
http://www.siam.org/journals/sima/33-6/37387.html

[†] Mathematics Department, 208 Clements Hall, Southern Methodist University, Dallas, TX 75275 (greddien@mail.smu.edu).

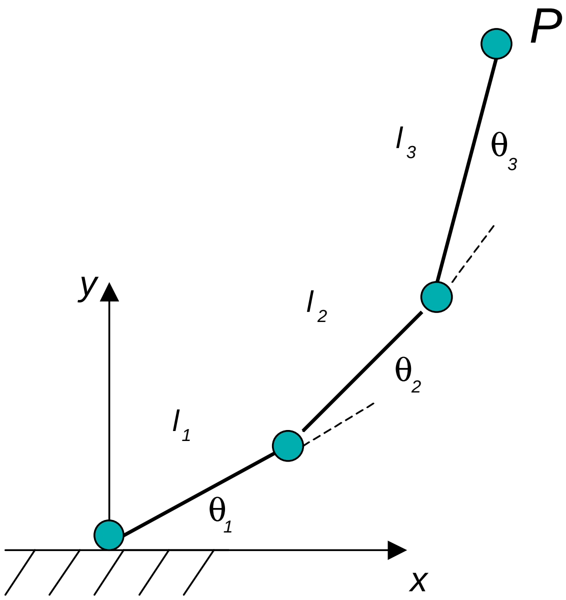


FIG. 1. *A planar manipulator.*

and define the singular set

$$S_1 = \{x \in D : \mathrm{rank}(F_x) = m - 1\}.$$

Then the boundary of $F(D)$ is contained in $F(S_1)$.

A basic computational problem in the study of mechanical devices such as the one represented by (1.1) is path-tracking, that is, to solve the equation

$$F(x(t)) = z(t) \tag{1.4}$$

for $x(t)$, where $z(t)$ is a given, parametrized path in the reachable set $F(D)$. At points where $F_x(x(t))$ has full rank, there is in principle no problem in solving (1.4) for a smooth path $x(t)$. But computational problems arise if $x(t)$ goes through a singularity. For the three-bar mechanism in Figure 1, let the motion be unconstrained. Then singularities exist for $\theta_3 = 0$ or π, for example, configurations likely to be encountered. The difficulties include possible lack of existence of solutions to (1.4) or existence without smoothness in the parameter t.

The purpose of this paper is to consider path-tracking through singularities. We use a unified framework for computing singularities (see [6]) which leads to efficient characterizations of the singular boundary and, as we show, to the criteria needed to establish the existence and smoothness of paths through singularities. Our primary focus is on the fold and the cusp cases. The method given for numerically computing and analyzing points where F has a rank drop is based on an implicit Liapunov–Schmidt reduction and has been shown to be useful in many applications. See, for example, [6], [7], [8], [9], and [16]. In section 2, we show how this reduction can be used to compute efficiently the various quantities that are needed in our analysis. In section 3, we use the formulas of section 2 to study the path-tracking problem in the smooth case. We relate path-tracking to classical results in singularity theory in section 4 and show how the results of section 3 can be modified to treat two basic nonsmooth cases.

2. Reduction. In this section we present a scalar function developed in [7], [8] that can be used to locate and characterize singularities. Functions that serve this purpose are often called test functions [17], and a natural choice is $\det(F_x)$. Indeed, this choice is used in the engineering literature [4], although the smallest singular value of F_x is also common [16]. Important advantages of the test function given below are its simple definition and evaluation, the existence of explicit derivative formulas with natural and inexpensive numerical approximations, and the fact that the approach provides a unified framework for turning points, bifurcation points, cusp points, and other higher order singularities and is applicable to maps F from R^n into R^m with $n > m$. These different cases generally entail ad hoc approaches in the engineering literature. These advantages are discussed further below.

Choose $r \in R^m$ and $T \in R^{p\times n}$ with $p = n - m + 1$ so that the bordered matrix

$$A = \begin{pmatrix} F_x & r \\ T & 0 \end{pmatrix}$$

is nonsingular on a neighborhood of a point x_0 in D. Consider the system of $(n+1)$-equations in the variables x, $f \in R$, and $s \in R^p$ given by

$$\begin{aligned} F - fr &= F(x_0), \\ Tx &= s + Tx_0. \end{aligned} \tag{2.1}$$

Note that $x = x_0$, $f = 0$, and $s_0 = 0$ is a solution to (2.1). The linearization of (2.1) in x and f at this point gives the matrix A (solving for $-f$), and so by the implicit function theorem, (2.1) defines $f = f(s)$ and $x = x(s)$ for s in a neighborhood of $s = 0$. We also note the following two systems for finding f_s, which is a row vector and equals the transpose of the gradient of f in s, and $H(f)$, the corresponding Hessian of f, which results from differentiating (2.1):

$$\begin{aligned} &F_x x_s - r f_s = 0, \quad && F_x x_{s_i s} - r f_{s_i s} = -F_{xx} x_{s_i} x_s, \\ &T x_s = I_p, && T x_{s_i s} = \ 0, \end{aligned} \tag{2.2}$$

$i = 1, \ldots, p$, where I_p is the $p \times p$ identity matrix. As in [6] and [7], define $u \in R^m$ to be the solution to (an exponent t denotes transpose, not time)

$$\begin{aligned} u^t F_x - f_s T &= 0, \\ u^t r &= 1. \end{aligned} \tag{2.3}$$

An easy calculation [7] shows that f_s may be used in both (2.2) and (2.3) and moreover that

$$f_s = u^t F_x x_s. \tag{2.4}$$

Writing $x_s = (v_1, v_2, \ldots, v_p)$, then the Hessian of f, $H(f)$, is given by $H(f) = (h_{ij})$ where $h_{ij} = u^t F_{xx} v_i v_j$. The Hessian matrix H is $p \times p$ and, of course, symmetric. The scalar function f is known in the literature as an implicit Liapunov–Schmidt reduction of F near x_0. This reduction has several advantages, both conceptually and computationally, over a standard Liapunov–Schmidt reduction [5]. The clearest and most important distinction is that the reduction does not depend on the a priori location of a point x_0 for which F_x has a rank drop. The systems (2.2) are solvable even if $F_x(x_0)$ has rank $m-1$. If $F_x^0 = F_x(x_0)$ has a rank drop, then from (2.4), $u_0^t F_x^0 = 0$ and so $f_s(0) = 0$. Points where the mapping F has a rank drop correspond to singular points for the reduced scalar function f.

Note that f defined in (2.1) depends on the vector r. The first equation in (2.1) can be written as $F(x) = F(x_0) + fr$, and so solutions to (2.1) are points x so that F hits points in the range moving along r from $F(x_0)$. In the case of (1.1), we could choose r to be the one-parameter family of radius vectors from the origin defined by the angle r makes with the abscissa in R^2. From (2.3) we can find points in the singular set, S_1, by solving the equation

$$f_s = 0. \tag{2.5}$$

The system (2.5) is square in general: p-equations in p-unknowns. In the case of (1.1) it represents two equations in two unknowns; the choice of r makes (2.5) have an extra variable, and so r can be varied and S_1 traced out.

This reduction can be used to determine whether a point $z_0 = F(x_0)$ is a boundary point. The following proposition can be found in [20]. Minimal smoothness assumptions on F are given there. In order to simplify the presentation given here, we will simply assume throughout that F is smooth. Recall the earlier assumption (H1) that $\operatorname{rank}(F_x) \geq m-1$ on D.

Theorem 2.1 (Griewank). *Let* (H1) *hold. Then the mapping F is open at x_0 if and only if the reduced function f defined in* (2.1) *is open at the origin.*

If F is not open at x_0, then f will not be open at the origin. Since f is a scalar function, then f must have an extremum at the origin, and f will not be open if it has either a maximum or a minimum. This can be checked in most cases by determining if the Hessian of f is negative or positive definite. In the case that H is semidefinite, higher order criteria must be used to determine if f has a maximum or a minimum (or neither). Then for $x_0 \in S_1, F(x_0)$ is a candidate to be a boundary point. At $x_0, f_s(0) = 0$. One can then compute H from (2.2). If H is indefinite, then $F(x_0)$ is not a boundary point. If H is either positive or negative definite, $F(x_0)$ is a local boundary point. The mechanism is then constrained locally, that is, $F(x_0)$ is a barrier point. Other points near $F(x_0)$ may be reachable, but they are then the image of points x not near x_0. Finally, if H is semidefinite, a higher order expansion and further analysis are necessary. We summarize these conclusions in the next theorem.

Theorem 2.2. *Let* (H1) *hold and let $x_0 \in S_1$. Then f defined by* (2.1) *satisfies $f_s(0) = 0$. If H_0 is negative definite, f has a maximum at $s = 0$ and x_0 is the preimage of a local boundary point. If H_0 is positive definite, then f has a minimum at $s = 0$ and x_0 is the preimage of a local boundary point. If H_0 is indefinite, then $F(x_0)$ is not a boundary point.*

Restricting our attention to the case where if H is indefinite, then $\lambda = 0$ is a simple eigenvalue for H, the higher order cases can then be analyzed as follows. Let f be given from (2.1) with $x_0 \in S_1$ so that $f_s^0 = 0$. We assume $Hc_p = 0$ with $c_p^t c_p = 1$. Let $\{c_1, c_2, \ldots, c_p\}$ be a basis for R^p so that $s \in R^p$ may be written as $s = Ct$ where $C = (c_i)$. Define $\hat{f}(t) = f(s)$. Now it follows that $\hat{f}_t(0) = 0$ since $f_s(0) = 0$ and that $\hat{f}_{t_i t_j}(0) = c_i^t H c_j$. We have assumed that $Hc_p = 0$ and $Hc_i \neq 0$, $i = 1, \ldots, p-1$. Then it follows that $\hat{f}_{t_i t_p}(0) = c_i^t H c_p = 0$ and that by assumption the $(p-1) \times (p-1)$ submatrix $D = (\hat{f}_{t_i t_j})_{i=1,j=1}^{p-1,p-1}$ of $H(\hat{f})$ is nonsingular. Using the decomposition lemma of singularity theory [14], $\hat{f}$ is isomorphic to a germ of the form $Q(t_1, \ldots, t_p) + \phi(t_p)$

with $\phi \in M^3$ ($M^3 = M \cdot M \cdot M$ where M is the ideal of germs equal to zero at the origin) and Q is a Morse germ. Based on this form, one can classify the cases that occur when $\hat{f}$ has low codimension. In fact, every germ of corank 1 and of codimension $\nu \geq 2$ is isomorphic to one and only one of the following germs:

$$\begin{aligned} &\textstyle\sum_{i=1}^{p-1} \epsilon_i t_i^2 + t_p^{\nu+1}, \quad \epsilon_i = \pm 1, \quad \text{if } \nu \text{ is even,} \\ &\textstyle\sum_{i=1}^{p-1} \epsilon_i t_i^2 \pm t_p^{\nu+1}, \quad \epsilon_i = \pm 1, \quad \text{if } \nu \text{ is odd.} \end{aligned} \tag{2.6}$$

In the case where H is semidefinite, it is then possible using (2.6) to develop a corresponding classification for boundary points. First note from (2.6) that if ν is even, then x_0 cannot be a boundary point because of the odd-order monomial in (2.6). If ν is odd, then depending on the sign of the $t_p^{\nu+1}$-term, $F(x_0)$ may or may not be a boundary point. In order to determine whether or not $F(x_0)$ is a boundary point in the semidefinite case, one must translate the criteria from ϕ back to f, which is a computable function. We do not do this here but refer the reader to [20].

Since there is no explicit s-dependence in (2.2), we define $g(x) = f_s$ and $x_s = V \in R^{nxp}$. It also follows from (2.2) that $g_x = u^t F_{xx} V$. This notation is consistent with [6]. The set S_1 is defined by the solutions to $g(x) = 0$. Generically, one expects S_1 to be a manifold because g should have full rank. But in problems such as (1.1), this is not the case at many points. Indeed, in mechanical design problems, including (1.1), $F(S_1)$ can be complicated, with multiple bifurcation points. See the analysis of the planar manipulator in [10], for example, or the Stewart platform [13]. The analysis of bifurcations can be included in the framework here.

3. Smooth paths. We next use the mathematical framework of section 2 to analyze the path-tracking problem: Given a path $z(t)$ in the work space, find a smooth path $x(t)$ in the domain of the manipulator that satisfies the relation

$$F(x(t)) = z(t) \tag{3.1}$$

for t in some interval. The analysis to follow puts some of the results of [11], [12], [15], [17], and [18] into our framework and leads to a new form for solvability conditions and computational algorithms. With $F: R^n \to R^m$ and $n > m$, there should be several solution paths $x(t)$. Theorem 2.2 can be used to establish the existence of a path in the cases where F has a simple rank drop. Assuming for the moment the existence and differentiability of $x(t)$ solving (3.1), we differentiate and obtain the system of differential equations

$$F_x(x(t)) \frac{dx}{dt} = \frac{dz}{dt} \tag{3.2}$$

with the initial condition $F(x_0) = z_0$. In this section we develop conditions within our framework to guarantee that a solution path $x(t)$ with continuous and bounded derivatives exists through points in S_1.

Let $x(t_0) = x_0 = 0$ with $x_0 \in S_1$. We assume, as before, that (H1) holds for F in a neighborhood of x_0. From (2.4) we immediately find the consistency condition

$$u_0^t \dot{z}_0 = 0, \tag{3.3}$$

where we assume $\dot{z}_0 = \frac{dz}{dt}(t_0) \neq 0$. Extend u_0 to form a basis $u_0, u_1, \ldots, u_{m-1}$ for R^m. Then the system of equations (3.2) can be put into the form

$$\begin{aligned} &\text{(a)} \quad U_{m-1}^t F_x(x(t)) \dot{x}(t) = U_{m-1}^t \dot{z}(t), \\ &\text{(b)} \quad u_0^t F_x(x(t)) \dot{x}(t) = u_0^t \dot{z}(t), \end{aligned} \tag{3.4}$$

where $U_{m-1} = (u_1, \ldots, u_{m-1})$. Assuming the needed smoothness, we differentiate (3.2) to obtain the equation

$$F_x \ddot{x} + F_{xx}\dot{x}\dot{x} = \ddot{z}, \tag{3.5}$$

from which it follows that

$$u_0^t F_x \ddot{x} + u_0^t F_{xx}\dot{x}\dot{x} = u_0^t \ddot{z}. \tag{3.6}$$

We find it convenient to make the following change of variables: Let $x(t) = t\,y(t)$ so that $\dot{x} = y + t\dot{y}$ where, without loss in generality, we let $t_0 = 0$ and $x_0 = 0$ so that $\dot{x}_0 = y_0$. Then (3.4) and (3.6) may be written at $t_0 = 0$ in the form

$$\begin{aligned} &\text{(a)} \quad U_{m-1}^t F_x^0 y_0 = U_{m-1}^t \dot{z}_0, \\ &\text{(b)} \quad u_0^t \dot{z}_0 = 0, \\ &\text{(c)} \quad u_0^t F_{xx}^0 y_0 y_0 = u_0^t \ddot{z}_0. \end{aligned} \tag{3.7}$$

The original equation (3.1) may be similarly decomposed as

$$\begin{aligned} &\text{(a)} \quad U_{m-1}^t F(x(t)) = U_{m-1}^t z(t), \\ &\text{(b)} \quad u_0^t F(x(t)) = u_0^t z(t). \end{aligned} \tag{3.8}$$

Now in order to show solvability and then differentiability of solutions to (3.1), we must first show that y_0 solving (3.7) exists. This system is not necessarily solvable for any path $z(t)$. Since $u_0^t \dot{z}_0 = 0$, we have that $\dot{z}_0$ is in the range of F_x^0, and so we may solve $F_x^0 y = \dot{z}_0$ for $y_0 = y_1 + V_0\alpha$, where the columns of V_0 span the null space of F_x^0, $\alpha \in R^p$ is arbitrary, and $F_x^0 y_1 = \dot{z}_0$ for a chosen particular solution y_1. For definiteness, let $Ty_1 = 0$. Then clearly (3.7a)–(3.7b) are satisfied. Substituting into (3.7c) we obtain

$$u_0^t F_{xx}^0 (y_1 + V_0\alpha)(y_1 + V_0\alpha) = u_0^t \ddot{z}_0 \tag{3.9}$$

or

$$\alpha^t (u_0^t F_{xx}^0 V_0 V_0)\alpha + 2\, u_0^t F_{xx}^0 (V_0\alpha) y_1 + u_0^t F_{xx}^0 y_1 y_1 = u_0^t \ddot{z}_0\,. \tag{3.10}$$

Equation (3.10) is a quadratic in α; if $u_0^t F_{xx}^0 V_0 V_0 = H$ is positive or negative definite, then (3.10) has a minimum or a maximum, respectively, at $\hat{\alpha}$ satisfying $u_0^t F_{xx}^0 V_0 V_0 \hat{\alpha} = -u_0^t F_{xx}^0 V_0 y_1$. Then depending on the sign and magnitude of $u_0^t \ddot{z}_0$, (3.10) can be solved. We note the following result.

THEOREM 3.1. *Let $x_0 \in S_1$ and let $H_0 = u_0^t F_{xx}^0 V_0 V_0 = g_x^0 V_0$ be indefinite. Then* (3.7) *is solvable.*

Theorem 3.1 is consistent with the results of section 2. If H is definite, motion from x_0 is possible in only one direction along r from $z_0 = F(x_0)$. Motion is possible in any direction when H is indefinite. The case in which H is nonsingular is called the fold.

Still assuming the existence and smoothness of a path $x(t)$, Taylor's formula with remainder allows us to write F and z in the following two forms:

$$\begin{aligned} &\text{(a)} \quad F(x(t)) = F^0 + R_1(x)x(t), \\ &\text{(b)} \quad F(x(t)) = F^0 + F_x^0 x(t) + R_2(x)x(t)x(t), \\ &\text{(c)} \quad z(t) = z_0 + \hat{R}_1(\dot{z})t, \\ &\text{(d)} \quad z(t) = z_0 + \dot{z}_0 t + \hat{R}_2(\ddot{z})t^2. \end{aligned} \tag{3.11}$$

These equations simplify under our normalization conditions that $x_0 = 0$ and $F(x_0) = z_0 = 0$. Using (3.11a), (3.11c) in (3.8a) and (3.11b), (3.11d) in (3.8b) with $x(t) = ty(t)$, we obtain

$$\text{(3.12)} \qquad \begin{aligned} &\text{(a)} \quad U_{m-1}^t R_1(ty(t))y(t) = U_{m-1}^t \hat{R}_1(\dot{z}(t)), \\ &\text{(b)} \quad u_0^t R_2(ty(t))y(t)y(t) = u_0^t \hat{R}_2(\ddot{z}(t)) \end{aligned}$$

as equations equivalent to (3.1) and (3.8). Note that for $t = t_0 = 0$, the system (3.12) becomes (3.7). The system (3.12) has the solution $y = y_0$ at $t = 0$. We consider (3.12) as m-equations in the n-unknowns y and with t as a parameter. We next show by an application of the implicit function theorem that (3.12) is solvable under the assumptions of Theorem 3.1.

THEOREM 3.2. *Let H_0 be indefinite and $z(t)$ have three continuous derivatives in a neighborhood of t_0. Then* (3.12) *is solvable for $y = y(t)$ with y continuously differentiable in a neighborhood of t_0. Moreover, $x = ty$ is continuously differentiable and solves* (3.1). *Since* (3.12) *is $m \times n$ plus the parameter t, $n - m$ of the coordinates in y are arbitrary.*

Proof. Since H_0 is indefinite, (3.10) has at least two distinct solutions, α_1 and α_2. Define $y_0^i = y_1 + V_0\alpha_i$ and suppose $g_x^0 y_0^i = 0$ for $i = 1, 2$. Subtracting, we obtain that $g_x^0 V_0(\alpha_1 - \alpha_2) = u_0^t F_{xx}^0 V_0 V_0(\alpha_1 - \alpha_2) = 0$, contradicting the indefiniteness of H_0. Thus there exists some y_0 solving (3.7) so that $g_x^0 y_0 \neq 0$. Since (3.7) is solvable, (3.12) is solvable at the origin in t with solution y_0. Linearizing (3.12) in y at $(y_0, 0)$, we obtain the homogeneous problem

$$\text{(3.13)} \qquad \begin{aligned} U_{m-1}^t (D_y R_1)|_{(y_0,0)} y_0 y + U_{m-1}^t F_x^0 y = 0, \\ u_0^t (D_y R_2)|_{(y_0,0)} y_0 y_0 y + \tfrac{1}{2} u_0^t F_{xx}^0 y_0 y = 0 \end{aligned}$$

to be solved for y. Since R_1 and R_2 contain the term ty, both derivatives in y are multiplied by t and so vanish at $t_0 = 0$. Thus (3.13) simplifies to

$$\text{(3.14)} \qquad \begin{aligned} &\text{(a)} \quad U_{m-1}^t F_x^0 y = 0, \\ &\text{(b)} \quad u_0^t F_{xx}^0 y_0 y = 0. \end{aligned}$$

The $(m \times n)$-dimensional system (3.14) has full rank in y at $t_0 = 0$ if the number of independent null vectors y is $n - m$. Equation (3.14a) has $p = n - m + 1$ solutions, namely V_0. Writing $y = V_0\alpha$ and substituting into (3.14b) give the equation

$$\text{(3.15)} \qquad \alpha^t (u_0^t F_{xx}^0 V_0 y_0) = 0,$$

or $\alpha^t (g_x^0 y_0) = 0$. Since $g_x^0 y_0 \neq 0$, (3.15) can have at most $(p-1)$ independent solutions α. Then y lies in a $(p - 1 = n - m)$-dimensional space, giving the implicit function theorem and the existence of $y(t)$. The smoothness assertion also follows from the implicit function theorem.

Remarks. If a solution to (3.10) exists satisfying the condition $g_x^0 y_0 \neq 0$ and if also the necessary condition (3.3) holds, then a smooth path of solutions to (3.1) exists. Theorem 3.2 gives a simple condition for this to happen. If the condition $g_x^0 y_0 \neq 0$ fails to hold, then $\dot{x}_0$ is tangent to S_1. Define $\mathcal{F}(y, t)$ to be the equations given in (3.12) written in the form $\mathcal{F}(y, t) = 0$. Left null vectors for $\mathcal{F}_y^0$ of the form $(\xi^t \, \nu)^t$ must solve

$$\text{(3.16)} \qquad \xi^t U_{m-1}^t F_x^0 + \nu u_0^t F_{xx}^0 y_0 = 0.$$

By the Fredholm alternative, (3.16) is solvable with nonzero ν if and only if $u_0^t F_{xx}^0 V_0 y_0 = g_x^0 y_0 = 0$. If $g_x^0 y_0 \neq 0$, then $\nu = 0$ and $\xi = 0$ since u_0 is assumed to be the only left null vector. Thus we see that $g_x^0 y_0 \neq 0$ implies $\mathcal{F}_y^0$ has full rank, giving a slightly different proof of Theorem 3.2. The condition that $\dot{x}_0$ is tangent to S_1 corresponds to a rank drop for (3.13). This means that $\mathcal{F}$ likely has a turning point or bifurcation point, and the existence of solutions is still possible. We do not pursue these cases here.

We next consider briefly the special case of self-motions [15]: These are paths $x(t)$ for which $F(x(t))$ is fixed. In the unconstrained three-bar mechanism of Figure 1, let $l_2 = l_3$. Then a self-motion obtains by holding θ_1 fixed and $\theta_3 = \pi$ and letting θ_2 vary. For redundant manipulators $(n > m)$, it is possible to use a self-motion to escape a singularity, which can be important in applications. We assume $z(t) = z_0$ with x_0 a singularity for F. Equation (3.3) is automatically satisfied and from (3.4) it follows that $\dot{x}_0$ is in the null space of F_x^0; that is, $y_0 = \dot{x}_0 = V_0\alpha$. Then (3.7c) becomes $u_0^t F_{xx}^0 y_0 y_0 = \alpha^t (u_0^t F_{xx}^0 V_0 V_0)\alpha = \alpha^t H \alpha = u_0^t \ddot{z}_0$. Solutions to this equation with $\alpha \neq 0$ require H to be indefinite. We then have the next theorem.

Theorem 3.3. *Using the definitions for self-motion through a singularity given above, let H be indefinite at x_0. Then there exists a nontrivial $y_0 = V_0\alpha$ satisfying $u_0^t F_{xx}^0 y_0 y_0 = 0$. In addition, there exists a solution $x(t)$ to* (3.1) *with two continuous derivatives and $\dot{x}_0 = y_0$.*

Proof. Write $x(t) = ty(t)$ as in the proof of Theorem 3.2. Since H is indefinite, then, in particular, H is nonsingular and $H\alpha \neq 0$ where the existence of $y_0 = V_0\alpha$ satisfying $\alpha^t H \alpha = 0$ was established above. This means that $g_x^0 V_0 \alpha \neq 0$ and so y_0 is not tangent to the singular manifold S_1. But then arguing exactly as in the proof of Theorem 3.2, the result follows.

Returning to the example above, the matrix H_0 has eigenvalues $-1 \pm 2\sqrt{2}$, and so the conditions of Theorem 3.3 are met. However, the self-motion is not the rotation described above but rather is one that has x_t^0 proportional to $(\,2\ 1\ 2)^t$. This leads to a self-motion that takes the robot to a nonsingular configuration. The rotation is a motion within the singular manifold.

Finally, we consider the case that the matrix H has a one-dimensional null space corresponding to $\nu = 2$ in (2.6). We add a nondegeneracy condition and establish the existence of a smooth path of solutions in the next theorem in the same manner as Theorem 3.2. Existence of a path also follows from Theorem 2.2 and (2.6). The conditions mean that F has a cusp singularity; details are given in section 4.

Theorem 3.4. *Let H_0 have a simple rank drop with null vector α_0 so that $u_0^t F_{xx}^0 v_0 v_0 = 0$ where $v_0 = V_0\alpha_0$. Define y_1 as the solution to $F_x^0 y_1 = \dot{z}_0$ with $Ty_1 = 0$, where z satisfies* (3.3) *and $\dot{z}_0 \neq 0$. In addition, let $u_0^t F_{xx}^0 v_0 y_1 \neq 0$. Then* (3.12) *is solvable for $y = y(t)$ with y continuously differentiable in a neighborhood of t_0. Moreover, $x = ty$ is continuously differentiable and solves* (3.1)*. Since* (3.12) *is $m \times n$ plus the parameter t, $n - m$ of the coordinates in y are arbitrary.*

Proof. As before, we write $x(t)$ in the form $x = ty$. We first note that by choosing α in (3.10) to be a scalar times α_0 where $v_0 = V_0\alpha_0$, then (3.10) is uniquely solvable. The quadratic term drops out with this choice. Arguing as in the proof of Theorem 3.2, the result follows if (3.14) has at most $(p-1)$ independent solutions. From (3.14a) it follows that $y = V_0\beta$. Substituting into (3.14b) gives the equation $\beta^t(u_0^t F_{xx}^0 V_0 y_0) = 0$, as before. Writing $y_0 = y_1 + V_0\alpha_0$, we obtain the equivalent equation $\beta^t(u_0^t F_{xx}^0 V_0 y_1) = 0$. But this equation is not satisfied if $\beta = \alpha_0$, and so (3.14) has $p - 1 = n - m$ independent solutions, completing the proof.

4. Nonsmooth solution paths.

In this section we study further the singularities in section 3 and relate the path-tracking problem for them to some classical results in singularity theory. Existence of solution paths in Theorem 3.2, the fold singularity, is problematical, but not in the case of Theorem 3.4, the cusp singularity. Conditions for the existence of smooth solutions were given in section 3. Singularity theory provides normal forms for these mappings that can be used to analyze their qualitative behavior. In particular, we show what can be done if smoothness is lost, which happens when condition (3.3) does not hold but paths still exist.

4.1. Singularity theory.

Two mappings F_1 and F_2 from manifolds X into Y are said to be equivalent if there exist diffeomorphisms $h : X \mapsto X$ and $k : Y \mapsto Y$ such that $F_2 = k \circ F_1 \circ h^{-1}$. A smooth map F is said to be stable if every map sufficiently near to it (including a sufficiently large number of derivatives) is equivalent to it. These definitions can be considered local [1]. Equivalence sets up a one-to-one correspondence between inverse images and singularities for the two maps. So if F_2 has a simpler form, qualitative results about F_1 can be derived. It is difficult to find the diffeomorphisms h and k. Nonetheless, an algorithm has been proposed for path-tracking through fold singularities based on computing the diffeomorphisms directly [18].

The fold and cusp singularities can be developed more formally as follows. Recall that S_1 denotes the set of points in X where F has a simple rank drop and that we have assumed F to have no other singularities. We need the regularity assumption that S_1 is a smooth manifold. Then $S_{1,1}$ is defined to be the points where F restricted to S_1 has a simple rank drop. Assuming $S_{1,1}$ is a smooth manifold, this definition may be continued.

We first consider the fold case. That is, we assume (H2) S_1 contains x_0, S_1 is a locally smooth manifold in a neighborhood of x_0, and $T_{x_0}S_1(F) + Null(F_x(x_0)) = T_{x_0}X$. These conditions guarantee that H is nonsingular. We next give the normal form in the equidimensional case.

Theorem 4.1. *Let F be a smooth mapping satisfying the conditions* (H2) *at x_0. Then there exist coordinates x centered at x_0 and coordinates y centered at $F(x_0)$ so that F has the form*

$$F : (x_1, \ldots, x_n) \mapsto (x_1, \ldots, x_{n-1}, x_n^2).$$

This theorem can be found in more generality in [5]. In effect, it means that diffeomorphisms h and k exist so that $k \circ F \circ h^{-1}$ has the above form. Moreover, it is also shown in [5] that this singularity is stable. See also [19].

Based on Theorem 4.1, more insight can be gained into the case considered in section 3. Consider, for example, the mapping $F : (x\ y) \mapsto (x\ y^2)$ from R^2 into R^2, which has a fold singularity at the origin and is in normal form. Note that $u_0 = \mathbf{e_2}$ spans $null(F_x^0)$ and $\mathbf{e_1}$ spans range(F_x^0). Let $z(t) = t(1\ \ 1)^t$, for example. Then $u_0^t \dot{z}(0) \neq 0$, which violates the basic condition (3.3). After eliminating t, define $U_\delta = \{z : z_1 - z_2 = 0,\, z = (z_1\ z_2)^t\}$, which is the set of points on z; it can be shown [3] that these conditions guarantee that locally $F^{-1}(U_\delta) = (x\ y)^t : x - y^2 = 0$ is a smooth manifold. But in terms of the parameter t, $t \geq 0$, we have $x = t\,,\ y = \pm t^{1/2}$, and smoothness does not obtain in t. This behavior is generic for the fold singularity if $u_0^t \dot{z}(0) \neq 0$. With $u_0^t \dot{z}_0 = 0$ in the above example, then $z(t) = c_1 t\mathbf{e}_1 + \frac{1}{2}t^2 \ddot{z}_0 + \mathcal{O}(t^3)$. Under F^{-1}, solution paths are $x(t) = c_1 t + \mathcal{O}(t^2)$ and $y(t) = \pm c_2 t + \mathcal{O}(t^2)$ provided that $c_2 = u_0^t \ddot{z}_0 > 0$. Note that this condition means that $\dot{x}_0$ is not tangent to S_1, which is the x-axis. One can compute (3.10) in this case and derive the equation

$2\alpha^2 = u_0^t \ddot{z}_0$. Two branches in the domain then map onto $z(t)$. If $c_2 < 0$, there are no solutions. These results are consistent with Theorem 3.2. Since $m = n$ in this case, one expects uniqueness from Theorem 3.2. Uniqueness does obtain, but locally about the two different solutions found for $y_0 = \dot{x}_0$ based on the different values for α in (3.10).

We next consider the cusp case. We assume (H3) $S_{1,1}$ contains x_0, S_1 and $S_{1,1}$ are smooth manifolds in a neighborhood of x_0, and F restricted to $S_{1,1}$ has full rank. The condition that $S_{1,1}$ be a smooth manifold through x_0 is satisfied if the system

$$g = 0, \ \det(H) = 0 \tag{4.1}$$

has full rank. The condition that F, restricted to $S_{1,1}$, has full rank is satisfied if the classical cusp condition holds; see (4.11) and [2] or [8]. In this case (H3) implies that $\nu = 2$ in (2.6). We then have the following theorem.

THEOREM 4.2. *Let F be a smooth mapping satisfying the conditions* (H3) *at x_0. Then there exist coordinates x centered at x_0 and coordinates y centered at $F(x_0)$ so that F has the form*

$$F : (x_1, \dots, x_n) \mapsto (x_1, \dots, x_{n-1}, x_{n-1}x_n - x_n^3).$$

The mapping in Theorem 4.2 is also called a generalized Whitney cusp and is stable. These singularities and normal forms include cases $n > m$ and higher order singularities in the pattern of Theorems 4.1 and 4.2. See [1] for general statements of these results. Special cases are included in [5] and [15]. Based on Theorem 4.2 and the general solvability of cubic equations, it follows that solution paths exist through the Whitney cusp. This analysis is an alternative to the approach outlined in section 2. Uniqueness and smoothness are the central issues. For example, let the mapping F from R^2 into R^2 be given by

$$F : (x\,y) \mapsto (x \quad xy - y^3), \tag{4.2}$$

which is the normal form for the cusp. The manifold S_1 for F can be parametrized as $(3t^2,\ t)^t$ with image under F given by $F(S_1) = (3t^2,\ 2t^3)^t$. The curve $F(S_1)$ defines a cusp at the origin centered about the positive x-axis. Within the region defined by this cusp curve, every point has three preimages under F. Outside of the cusp region, preimages are unique. Thus paths $x(t)$ exist for any $z(t)$. Consider the example $z(t) = (0,\ t)^t$. One can solve explicitly the equation $F(x(t)) = z(t)$ obtaining the curve $x(t) = (0\,,\ -t^{1/3})^t$, showing that smoothness is lost. In this case, $u_0^t \dot{z}(0) \neq 0$. This is the generic behavior when (3.3) does not hold.

The lack of smoothness in the computed paths in the above examples is due to the parametrizations. We next show how these cases can be treated. The approach is to find an appropriate parametrization for which a smooth expansion is possible using the classical method of Puiseux series.

4.2. Puiseux series. Let $x = x(s)$ and $t = t(s)$, where s is a new parameter. We let s be arclength by requiring $x_s(s)^t x_s(s) = 1$, where the subscript denotes differentiation in the indicated variable. Differentiating (3.1) in s we obtain

$$F_x x_s = z_t t_s. \tag{4.3}$$

If the solution path in s exists and is smooth, then at the singularity (with $t_0 = s_0 = 0$), (4.3) holds. In the next two subsections we develop the first few terms in series expansions for both x and t in s following the proofs of Theorems 3.2 and 3.4. We highlight the differences in the proofs for these two cases when (3.3) does not hold and the reparametrization above is used.

Fold singularities. We assume that $u_0^t z_t^0 \neq 0$ since otherwise the results of section 3 apply. Then it must be the case that $t_s^0 = 0$. But then $x_s^0 = V_0\alpha$ with $\alpha^t V_0^t V_0 \alpha = 1$. As in section 3 we write $x(s) = sy(s)$, and we write $t(s) = s^2\tau(s)$. Since this is the fold case, we have $u_0^t F_{xx}^0 y_0 y_0 \neq 0$ where $x_s^0 = y_0$. Then with these notations, (3.7) becomes

$$u_0^t F_{xx}^0 y_0 y_0 = \alpha^t H \alpha = 2(u_0^t z_t^0)\tau_0. \tag{4.4}$$

Using τ_0 as an unknown, (4.4) can be solved given any α. We assume y_0 and τ_0 have been chosen satisfying (4.4) and $y_0^t y_0 = 1$. With an added approximate arclength normalization, the system (3.12) becomes in this case

$$\begin{gathered} U_{m-1}^t R_1(sy)y = \left(U_{m-1}^t \int_0^1 z_t(ts^2\tau)dt\right) s\tau, \\ u_0^t R_2(sy)yy = \left(u_0^t \int_0^1 z_t(ts^2\tau)dt\right)\tau, \\ y_0^t y = 1. \end{gathered} \tag{4.5}$$

By the above considerations, (4.5) is solvable at $s = 0$ and, when linearized at $s = 0$, gives the following linear system in the unknowns $\hat{y}, \hat{\tau}$:

$$\begin{gathered} U_{m-1}^t F_x^0 \hat{y} = 0, \\ \tfrac{1}{2}\, u_0^t F_{xx}^0 y_0 \hat{y} - (u_0^t z_t^0)\hat{\tau} = 0, \\ y_0^t \hat{y} = 0. \end{gathered} \tag{4.6}$$

From (4.6) we obtain that $\hat{y} = V_0\alpha$. But $y_0^t\hat{y} = 0$ forces α to lie in a $(p-1 = n-m)$-dimensional subspace of R^p. Since $u_0^t z_t^0 \neq 0$, $\hat{\tau}$ depends linearly on α. Thus (4.6) has $(n-m)$ independent solutions and (4.5) has full rank. Arguing as in Theorem 3.2, we obtain smooth solutions $y = y(s)$ and $\tau = \tau(s)$ of (4.5) if z is smooth, and hence of the original equations. Note, however, that if H is definite, then $\alpha^t H \alpha$ is of one sign independent of α. Assume, for example, that H is positive definite. If $u_0^t z_t^0$ is negative, then τ_0 must be negative in (4.4). However, with $t = s^2\tau$, this would mean that a solution has been found but with t negative. That is, a solution exists only backward in time, and, in effect, there is no solution. If H is indefinite, then α can be chosen so that τ_0 is positive.

Since $t(s) = s^2\tau(s)$ with $\tau_0 \neq 0$, then y depends on $t^{1/2}$, and the solution path sought based on the original parametrization of the target is not smooth. For a two-arm robot of the form (1.1) with $l_1 = l_2 = 1$, the equations in (1.1) become

$$\begin{gathered} z_1 = \cos(x) + \cos(x+y), \\ z_2 = \sin(x) + \sin(x+y). \end{gathered} \tag{4.7}$$

The origin is a singular configuration and it is easy to compute that for the mapping F defined by (4.7),

$$F_x^0 = \begin{pmatrix} 0 & 0 \\ 2 & 1 \end{pmatrix}, \quad F_{xx}^0 = \begin{pmatrix} -2 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{4.8}$$

S_1 is the x-axis, and $F(S_1)$ is the circle of radius 2 with center at the origin. We consider the path studied in [11], namely $z_1(t) = 2 - t$, $z_2(t) = 0$, for which

$z_t^0 = -\mathbf{e}_1$. Note that z_t^0 is not tangent to $F(S_1)$ at the origin in S_1. A right null vector for F_x^0 is $v_0 = (1 \ -2)^t$ and a left null vector is $u_0 = \mathbf{e}_1$. Since $u_0^t z_t^0 \neq 0$, we are in the above setting and the results of section 3 do not apply. Solving (4.5) using an arclength parametrization, we have $(x_{ss}^0)^t x_s^0 = 0$. Using (4.3), (4.4) we obtain $x_s^0 = (1/\sqrt{5}\ , \ -2/\sqrt{5})^t$ and $x_{ss}^0 = 0$. So $x(s) = x_s^0 s + \mathcal{O}(s^3)$ with $t = \frac{1}{5}s^2 + \cdots$. Thus $s = \sqrt{5} \cdot t^{\frac{1}{2}} + \cdots$ and $x(t) = (-t^{\frac{1}{2}}, \, 2t^{\frac{1}{2}})^t + \cdots$, consistent with the results of [12]. We note that $u_0^t F_{xx}^0 y_0 y_0 = -2$.

Cusp singularities. We next derive a quasi-arclength parametrizaton of a solution path through a cusp singularity. As before, we write $t = t(s)$ and solve $F(x(s)) = z(t(s))$. Assuming for the moment existence and smoothness of both x and t, we derive the conditions to be imposed as a modification to the proof of Theorem 3.4. Differentiating, we obtain as before (4.3). We consider only the case where $u_0^t z_t^0 \neq 0$, so that, as above, $t_s^0 = 0$ and $x_s^0 = V_0 \alpha$. Now in the cusp case, $H = u_0^t F_{xx}^0 V_0 V_0$ is a singular matrix. We choose $x_s^0 = \phi_0$ so that $u_0^t F_{xx}^0 \phi_0 \phi_0 = 0$, or equivalently, $H\alpha = 0$ with $\phi_0^t \phi_0 = 1$ as a normalization. Differentiating again we obtain the equation

$$F_x x_{ss} = -F_{xx} x_s x_s + z_{tt} t_s^2 + z_t t_{ss}. \tag{4.9}$$

Since $t_s^0 = 0$ and $u_0^t F_{xx}^0 \phi_0 \phi_0 = 0$, it follows that $t_{ss}^0 = 0$ and that (4.9) is solvable. We add the normalization condition $(x_s^0)^t x_{ss}^0 = 0$ for x_{ss}^0. Differentiating (4.9) we obtain

$$F_x x_{sss} = -3F_{xx} x_{ss} x_s - F_{xxx} x_s x_s x_s + z_{ttt} t_s^3 + z_t t_{sss} + 3 z_{tt} t_{ss} t_s, \tag{4.10}$$

from which it follows at s_0 that

$$3u_0^t F_{xx}^0 x_{ss}^0 x_s^0 + u_0^t F_{xxx}^0 x_s^0 x_s^0 x_s^0 = (u_0^t z_t^0) t_{sss}^0. \tag{4.11}$$

The left-hand side of (4.11) not being equal to zero is the standard cusp condition [7], under which it follows that $t_{sss}^0 \neq 0$.

Based on these preliminaries we set $x(s) = s\phi_0 + s^2 y$ and $t(s) = s^3 \tau$ with $\phi_0^t y = 0$. Then the system $F(x) = z$ is equivalent to

$$\begin{aligned} &\text{(a)} \quad U_{m-1}^t F(s\phi_0 + s^2 y) = U_{m-1}^t z(s^3 \tau), \\ &\text{(b)} \quad\quad u_0^t F(s\phi_0 + s^2 y) = \ u_0^t z(s^3 \tau), \\ &\text{(c)} \quad\quad\quad\quad\quad\quad \phi_0^t y \ = 0 \end{aligned} \tag{4.12}$$

with the preceding definitions. The system (4.12) consists of $(m+1)$-equations in the $(n+1)$-variables y, τ plus the parameter s; (4.12) has full rank at a point if the null space of its linearization there has dimension $n-m$. Using Taylor expansions, (4.12a) and (4.12b) may be written, respectively, as

$$\begin{aligned} &\text{(a}'\text{)} \quad U_{m-1}^t F_x^0 (s\phi_0 + s^2 y) + U_{m-1}^t R_2 (s\phi_0 + s^2 y)^2 = U_{m-1}^t \hat{R}_1 s^3 \tau, \\ &\text{(b}'\text{)} \quad u_0^t F_{xx}^0 (s\phi_0 + s^2 y)^2 + u_0^t R_3 (s\phi_0 + s^2 y)^3 = u_0^t \hat{R}_1 s^3 \tau. \end{aligned} \tag{4.12}$$

Using (4.12a$'$) and dividing out an s^2 and (4.12b$'$) and dividing out an s^3, the system (4.12) at $s = 0$ is equivalent to

$$\begin{gathered} U_{m-1}^t F_x^0 y + \tfrac{1}{2} U_{m-1}^t F_{xx}^0 \phi_0 \phi_0 = 0, \\ u_0^t F_{xx}^0 \phi_0 y + \tfrac{1}{6} u_0^t F_{xxx}^0 \phi_0 \phi_0 \phi_0 = (u_0^t z_t^0) \tau, \\ \phi_0^t y = 0, \end{gathered}$$

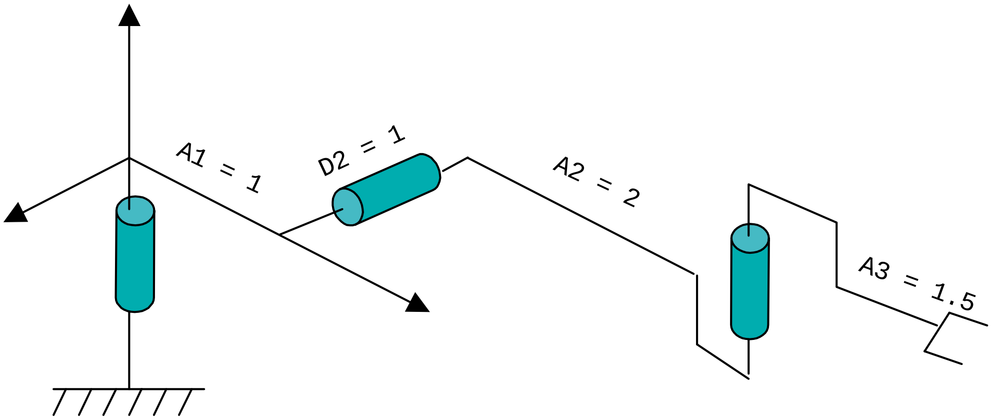


FIG. 2. *A spatial manipulator.*

which is solvable with $y_0 = \frac{1}{2}x^0_{ss}$ and $\tau_0 = t^0_{sss}/6$ as defined earlier. Linearizing (4.12) at y_0, τ_0, and s_0 produces the following linear system in $\hat{y}$, $\hat{\tau}$:

$$(4.13)\qquad \begin{aligned} &\text{(a)} \qquad U^t_{m-1}F^0_x\hat{y} = 0,\\ &\text{(b)} \qquad u^t_0F^0_{xx}\phi_0\hat{y} - (u^t_0z^0_t)\hat{\tau} = 0,\\ &\text{(c)} \qquad \phi^t_0\hat{y} = 0. \end{aligned}$$

From (4.13a) we have $\hat{y} = V_0\alpha$; from (4.13c) we have that α lies in a $(p-1)$-dimensional subspace of R^p. Now $u^t_0F^0_{xx}\phi_0\hat{y} = \alpha^tH\alpha_0 = 0$ for any α since ϕ_0 with α_0 a null vector for H. Then from (4.13b) it follows that $\hat{\tau} = 0$, and (4.13) admits $m - n = p - 1$ independent solutions. It follows using the implicit function theorem as in Theorem 3.4 that (4.12) is solvable for y and τ as smooth functions of s if z is smooth. Since $t = s^3\tau$, it follows that $s = t^{1/3} + \cdots$ and that $x = x^0_st^{1/3} + \frac{1}{2}x^0_{ss}t^{2/3} + \cdots$, which was expected from the discussion following (4.2).

As an example, we consider the three degree-of-freedom robot in Figure 2; see [4] and [13]. The motion of each joint is in the Lie group $SE(3)$, and the equations for the mapping from the three joint angles to the spatial coordinates of the end-effector may be simply computed as follows: Define the matrices

$$A = \begin{pmatrix} \cos\theta_1 & -\sin\theta_1 & 0 & 0\\ \sin\theta_1 & \cos\theta_1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} \cos\theta_2 & 0 & \sin\theta_2 & 1\\ 0 & 1 & 0 & 1\\ -\sin\theta_2 & 0 & \cos\theta_2 & 0\\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$C = \begin{pmatrix} \cos\theta_3 & -\sin\theta_3 & 0 & 2\\ \sin\theta_3 & \cos\theta_3 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad D = \begin{pmatrix} 1.5\\ 0\\ 0\\ 1 \end{pmatrix}.$$

The coordinate systems were taken with the positive x-axes along the links denoted by a_i, the positive z-axes vertical and upward, and with the separate systems right-handed. The first three coordinates of the product $ABCD$ give the mapping F. Based on this formula, one can easily compute F_θ and $F_{\theta\theta}$. Using the formulas given in section 2, it is straightforward to compute that F has a cusp point at $\theta_1 = 0$, $\theta_2 = -2.265134$, and $\theta_3 = 1.158639$. That is, for this point, F_θ has a one-dimensional null space, and the null vector v is tangent to the singular manifold S_1. This can be verified by computing $H = u^t F_{\theta\theta} vv = g_\theta v = 0$. The null vector, normalized by $T = (1\ 1\ 1)$ in (2.2), is $v^t = (0.356216\ 0.2499307\ 0.3938177)$; the left null vector u, normalized with $r^t = (1\ \ -1\ 1)$ as in (2.3), is $u^t = (0.2964187\ \ -1.0595738\ \ -0.3559925)$. Finally, using the differentiation formula $g_\theta = u^t F_{\theta\theta} v$, one computes $g_\theta = (0\ 0.4533926\ -0.2877052)$. We point out that these calculations are a by-product of solving the system $g = 0$, $H = 0$, to find the cusp point. Now all information is at hand to verify if a path $z(t)$ for the end-effector is the image of a smooth path in the space of angles. As discussed above and in Theorem 3.4, one needs $u^t \dot{z} = 0$ and then $g_\theta y_1 \neq 0$ for a smooth path where $F_\theta y_1 = \dot{z}$ at t_0. Otherwise, a reparametrization is required as discussed above.

Planar intersections of the image under F of the singular set can be computed by solving the system

$$\begin{aligned} g &= 0, \\ q^t F &= 0, \end{aligned} \tag{4.14}$$

where q is a fixed vector. We next give a proposition with conditions under which (4.14) is nonsingular at the singular point x_0 for F.

PROPOSITION 4.3. *Let* $q^t F_x^0 \neq 0$. (a) *Let* $u_0^t F_{xx}^0 \phi_0 \phi_0 \neq 0$. *Then the linearization of* (4.14) *has a one-dimensional null space at* x_0. (b) *Let* $u_0^t F_{xx}^0 \phi_0 \phi_0 = 0$. *Define* w *so that* $q^t w = 0$, $u_0^t w = 0$, *and* $w^t w = 1$. *Let* $F_x^0 v_1 = w$ *with* $w^t \phi_0 = 0$. *Then if* $u_0^t F_{xx}^0 \phi_0 v_1 \neq 0$, *the conclusion in* (a) *holds.*

Proof. For part (a), any null vector for the linearization of the second equation in (4.14) has the form $\alpha\phi_0 + \beta v_1$ where v_1 is defined above; then using the condition $u_0^t F_{xx}^0 \phi_0 \phi_0 \neq 0$ shows that α and β are linearly related. A similar substitution under the conditions of (b) shows that $\beta = 0$.

Under the conditions of (b) in Proposition 4.3 for this example, the equations (4.14) define a smooth curve in the singular surface. Since the tangent to this curve is by construction a null vector for F_x^0, the image of this curve will have a cusp.

Remarks. If the target path $z(t)$ satisfies $u_0^t \dot{z}_0 = 0$, then it was shown in section 3 that a smooth path of inverse images exists in the case of the cusp, and may exist in the case of the fold providing that the matrix H associated with the mapping F is indefinite. These results were generalized in section 4 to the case that $u_0^t \dot{z}_0 \neq 0$. Then the solutions are not smooth in the parameter t, but are smooth in a computable reparametrization. Conditions required for these results are the generic ones for the two singularities. Moreover, our results can be seen as an example of an effective method to analyze other singularities and path-tracking through them. One performs the analysis based on the normal form for the singularity and then relates the derived solvability conditions and expansions to the original problem through the coordinate changes relating the original problem to its normal form. These conditions can be evaluated using the reduction given in section 2. This approach is especially effective for questions related to the qualitative behavior of spatial manipulators near a singularity, such as determining the existence of a path and the number of inverse images.

REFERENCES

[1] V.I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, 1982.

[2] V.I. Arnold, S.M. Gusein-Zade, and A.N. Varchenko, *Singularities of Differential Maps*, Birkhäuser, Boston, 1985.

[3] T. Broecker and L. Lander, *Differentiable Germs and Catastrophes*, London Math. Soc. Lecture Note Ser. 17, Cambridge University Press, Cambridge, UK, 1975.

[4] C. Chevallereau, *Feasible trajectories in task space from a singularity for a nonredundant to redundant robot manipulator*, Internat. J. Robotics Res., 17 (1998), pp. 56–69.

[5] M. Golubitsky and V. Guillemin, *Stable Mappings and Their Singularities*, Springer-Verlag, New York, 1980.

[6] M. Golubitsky and D.G. Schaeffer, *Singularities and Groups in Bifurcation Theory*, Springer-Verlag, New York, 1985.

[7] A. Griewank and G.W. Reddien, *Characterization and computation of generalized turning points*, SIAM J. Numer. Anal., 21 (1984), pp. 176–185.

[8] A. Griewank and G.W. Reddien, *Computation of cusp singularities of operator equations and their discretizations*, J. Comput. Appl. Math., 26 (1989), pp. 133–153.

[9] A. Griewank and G.W. Reddien, *The approximate solution of defining equations for generalized turning points*, SIAM J. Numer. Anal., 33 (1996), pp. 1912–1920.

[10] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Appl. Math. Sci. 42, Springer-Verlag, New York, 1983.

[11] E.J. Haug, C.M. Luh, F.A. Adkins, and J.Y. Wang, *Numerical algorithms for mapping boundaries of manipulator workspaces*, Trans. ASME J. Mech. Des., 118 (1996), pp. 228–234.

[12] J. Kieffer, *Differential analysis of bifurcations and isolated singularities for robots and mechanisms*, IEEE Trans. Robotics Automat., 10 (1994), pp. 1–10.

[13] J.E. Lloyd, *Desingularization of nonredundant serial manipulator trajectories using Puiseux series*, IEEE Trans. Robotics Automat., 14 (1998), pp. 590–600.

[14] C.M. Luh, F.A. Adkins, E.J. Haug, and C.C. Qiu, *Working capability analysis of Stewart platforms*, Trans. ASME J. Mech. Des., 118 (1996), pp. 220–227.

[15] J. Martinet, *Singularities of Smooth Functions and Maps*, London Math. Soc. Lecture Note Ser. 58, Cambridge University Press, Cambridge, UK, 1982.

[16] K.A. O'Neil, Y.C. Chen, and J. Seng, *Removing singularities of resolved motion control of mechanisms, including self-motion*, IEEE Trans. Robotics Automat., 13 (1997), pp. 741–751

[17] R. Seydel, *Practical Bifurcation and Stability Analysis: From Equilibrium to Chaos*, Springer-Verlag, New York, 1994.

[18] K. Tchon and R. Muszynski, *Singular inverse kinematic problem for robotic manipulators: A normal form approach*, IEEE Trans. Robotics Automat., 14 (1998), pp. 93–104.

[19] K. Tchon, *A normal form of singular kinematics of robot manipulator with smallest degeneracy*, IEEE Trans. Robotics Automat., 11 (1995), pp. 401–404.

[20] O. Vogel, *Zur Charakterisierung von Raendern und Randfaltungen von Vektorfunktionen der Klassen C^k und $CS^{1,1}$ mit Anwendung zur multikriterielle Optimierung*, Diplom Arbeit, Technische Universitaet Dresden, Dresden, Germany, 1998.

SIAM J. MATH. ANAL.
Vol. 33, No. 6, pp. 1320–1329

ZERO DIFFUSION-DISPERSION LIMITS FOR SCALAR CONSERVATION LAWS*

CEZAR I. KONDO[†] AND PHILIPPE G. LEFLOCH[‡]

Abstract. We consider solutions of hyperbolic conservation laws regularized with vanishing diffusion and dispersion terms. Following a pioneering work by Schonbek, we establish the convergence of the regularized solutions toward discontinuous solutions of the hyperbolic conservation law. The proof relies on the method of compensated compactness in the L^2 setting. Our result improves upon Schonbek's earlier results and provides an optimal condition on the balance between the relative sizes of the diffusion and the dispersion parameters. A convergence result is also established for multidimensional conservation laws by relying on DiPerna's uniqueness theorem for entropy measure-valued solutions.

Key words. conservation law, shock wave, entropy solution, measure-valued solution, diffusion, dispersion, singular limit, a priori estimate

AMS subject classifications. Primary, 35L65; Secondary, 76N10

PII. S0036141000374269

1. Introduction.

We study here the convergence of solutions of the partial differential equation ($\epsilon \to 0+$, $\delta = \delta(\epsilon) \to 0$)

$$u_t + f(u)_x = \epsilon\, u_{xx} + \delta\, u_{xxx}, \qquad u = u^\epsilon(x,t),\ x \in \mathbb{R},\ t \geq 0, \tag{1.1}$$

toward weak solutions of the corresponding hyperbolic conservation laws:

$$u_t + f(u)_x = 0, \qquad u = u(x,t),\ x \in \mathbb{R},\ t \geq 0, \tag{1.2}$$

where the flux $f : \mathbb{R} \to \mathbb{R}$ is a smooth function with (at most) linear growth at infinity; that is, for some $M > 0$

$$|f'(u)| \leq M, \qquad u \in \mathbb{R}.$$

Equations of the form (1.1)–(1.2) arise in fluid dynamics when both viscosity (diffusion) and capillarity (dispersion) play a role. The diffusion ϵ smoothes out the discontinuous solutions of (1.2), while the dispersion δ causes high-frequency oscillations.

In this paper, we establish that the solutions u^ϵ of (1.1) converge toward a weak solution of (1.2) provided that

$$\delta = O(\epsilon^2). \tag{1.3}$$

When the stronger condition

$$\delta = o(\epsilon^2) \tag{1.4}$$

[*] Received by the editors June 14, 2000; accepted for publication (in revised form) December 20, 2001; published electronically May 23, 2002.
http://www.siam.org/journals/sima/33-6/37426.html

[†] Universidade Federal de São Carlos, Departamento de Matemática, Caixa Postal 676, 13565-905, São Carlos-SP, Brazil (dcikdm.ufscar.br). This work was done when this author was visiting the Ecole Polytechnique during the academic year 1999–2000, with financial support from PICD-Capes (Brazil) and the European research project Alpha.

[‡] Centre de Mathématiques Appliquées and Centre National de la Recherche Scientifique, U.M.R. 7641, Ecole Polytechnique, 91128 Palaiseau, France (leflochcmap.polytechnique.fr).

holds, we prove that the limit coincides with the entropy solution determined by Kruzkov's theory [10]. We point out that these conditions are sharp since, in the limiting case

$$\delta = K\,\epsilon^2 \qquad \text{for some } K \in \mathbb{R}, \tag{1.5}$$

limiting solutions may violate Kruzkov's entropy conditions [8, 5, 15, 2]. Furthermore, when (1.3) is violated, the solutions are highly oscillatory and fail to converge in any strong topology as noted by Lax and Levermore [12, 13, 14]. (See also Lax [11].)

The singular limit problem above was first tackled by Schonbek [21], who established the optimal rate (1.3) for the Burgers equation, that is,

$$f(u) = \frac{u^2}{2},$$

and for the class of flux-functions

$$f(u) = \frac{u^{2p+1}}{2p+1}, \qquad p \geq 1.$$

She also gave a convergence result for general fluxes with quadratic growth at infinity, however, under the stronger condition on $\delta = O(\epsilon^3)$. As another important contribution in [21], Schonbek introduces a generalization of the method of compensated compactness (Tartar [23] and Murat [20]) allowing the handling of sequences that are bounded in L^p for finite $p > 1$ only. Next, following [21], LeFloch and Natalini [17] studied equations like (1.1) but with nonlinear (even singular) diffusion, and established strong convergence results toward entropy solutions of (1.2). See also a convergence result for systems in Hayes and LeFloch [6].

In the second part of this paper, we also deal with the convergence of solutions of multidimensional equations similar to (1.1)–(1.2). For multidimensional equations, the compensated compactness method no longer applies and the proofs are based instead on DiPerna's uniqueness theory for entropy measure-valued solutions (DiPerna [4], Szepessy [22], and Kondo and LeFloch [9]). Our approach is similar to Correia and LeFloch [3] where nonlinear diffusion terms are treated under a strong assumption on the ratio of the dispersion to the diffusion.

To summarize, the main contribution in the present paper is the derivation of a priori estimates (Theorems 2.1 and 3.1) which cover general flux-functions (with at most linear growth at infinity) and lead to an optimal condition on the balance between the diffusion and the dispersion (Theorems 2.2 and 3.2).

Further material on classical and nonclassical entropy solutions generated by diffusive-dispersive limits can be found in [1, 2, 5, 6, 7, 8, 15, 16, 17, 18, 19, 21]

2. One-dimensional conservation laws.

Consider a family u^ϵ of smooth solutions to

$$u_t + f(u)_x = \epsilon\, u_{xx} + \delta\, u_{xxx}, \qquad u = u^\epsilon(x,t), \tag{2.1}$$

$$u(x,0) = u_0^\epsilon(x), \qquad x \in \mathbb{R}, \tag{2.2}$$

where $\epsilon \to 0+$ and $\delta = \delta(\epsilon) \to 0$. Under suitable conditions on the initial data $u_0^\epsilon : \mathbb{R} \to \mathbb{R}$, the solutions (and their derivatives) decay rapidly at infinity so that all the a priori estimates given below are rigorously justified. We want to show that the solution of (2.1)–(2.2) converges toward a weak solution of the problem

$$u_t + f(u)_x = 0, \qquad u = u^\epsilon(x,t), \tag{2.3}$$

$$u(x,0) = u_0(x), \qquad x \in \mathbb{R}, \tag{2.4}$$

where $u_0 : \mathbb{R} \to \mathbb{R}$ are given initial data. A minimum requirement is the weak convergence (for instance in $L^2(\mathbb{R})$)

$$u_0^\epsilon \rightharpoonup u_0,$$

which is always assumed throughout this paper. The following convergence theorem covers both cases where the diffusion is in balance or dominates the dispersion.

Theorem 2.1. *Suppose that the flux-function f is Lipschitz continuous on $\mathbb{R}$ and that the initial data u_0 belong to $L^2(\mathbb{R})$. Then the solution u^ϵ of* (2.1)–(2.2) *satisfies the following a priori estimates:*

$$\|u^\epsilon(t)\|_{L^2(\mathbb{R})} \le \|u_0^\epsilon\|_{L^2(\mathbb{R})}, \qquad t \ge 0, \tag{2.5a}$$

$$\sqrt{2\,\epsilon}\, \|u_x^\epsilon\|_{L^1(\mathbb{R}^+,L^2(\mathbb{R}))} \le \|u_0^\epsilon\|_{L^2(\mathbb{R})}, \tag{2.5b}$$

$$\sqrt{\delta}\, \|u_x^\epsilon(t)\|_{L^2(\mathbb{R})} \le \sqrt{2\,\|f'\|_\infty}\, \|u_0^\epsilon\|_{L^2(\mathbb{R})} + \sqrt{\delta}\, \|u_{0x}^\epsilon\|_{L^2(\mathbb{R})}, \qquad t \ge 0, \tag{2.5c}$$

and

$$\sqrt{\epsilon\,\delta}\, \|u_{xx}\|_{L^1(\mathbb{R}^+,L^2(\mathbb{R}))} \le \sqrt{2\,\|f'\|_\infty}\, \|u_0^\epsilon\|_{L^2(\mathbb{R})} + \sqrt{\delta}\, \|u_{0x}^\epsilon\|_{L^2(\mathbb{R})}. \tag{2.5d}$$

Proof. Throughout the calculation and for simplicity, we omit the upper-index ϵ. To any smooth function $U : \mathbb{R} \to \mathbb{R}$ we can associate a “flux” $F : \mathbb{R} \to \mathbb{R}$ by $F'(u) = U'(u) f'(u)$, $u \in \mathbb{R}$. Multiplying (2.1) by $U'(u)$ we find

$$U(u)_t + F(u)_x = \epsilon\,(U'(u)u_x)_x - \epsilon\, U''(u)\, u_x^2 + \delta\,\big(U'(u)\, u_{xx}\big)_x - \delta U''(u)\, u_x\, u_{xx}.$$

Integrating over the whole space, it follows that

$$\begin{aligned} \frac{d}{dt}\int_{\mathbb{R}} U(u)\,dx + \epsilon \int_{\mathbb{R}} U''(u)\, u_x^2\,dx &= \delta \int_{\mathbb{R}} U''(u) \left(\frac{u_x^2}{2}\right)_x dx \\ &= -\frac{\delta}{2}\int_{\mathbb{R}} U'''(u)\, u_x^3\,dx. \end{aligned} \tag{2.6}$$

Integrating in time over some interval $(0,t)$, we arrive at the general identity

$$\int_{\mathbb{R}} U(u(t))\,dx + \epsilon \int_0^t \int_{\mathbb{R}} U''(u)\, u_x^2\,dxdt = \int_{\mathbb{R}} U(u_0)\,dx - \frac{\delta}{2}\int_0^t \int_{\mathbb{R}} U'''(u)\, u_x^3\,dxdt. \tag{2.7}$$

Choosing first $U(u) = u^2$ in (2.7), we see that

$$\int_{\mathbb{R}} u(t)^2 dx + 2\,\epsilon \int_0^t \int_{\mathbb{R}} u_x^2 dx = \int_{\mathbb{R}} u_0^2(x), \tag{2.8}$$

which gives immediately (2.5a) and (2.5b).

Next, we differentiate (2.1) with respect to x and we multiply by u_x:

$$\frac{1}{2}\left(u_x^2\right)_t + u_x\left(f'(u)\,u_x\right)_x = \epsilon\left(u_x\,u_{xx}\right)_x - \epsilon\,u_{xx}^2 + \delta\left(u_x\,u_{xxx} - \frac{1}{2}u_{xx}^2\right)_x.$$

Integrating in space, we get

$$\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}} u_x^2\,dx + \epsilon\int_{\mathbb{R}} u_{xx}^2\,dx = \int_{\mathbb{R}} u_{xx}\,f'(u)\,u_x\,dx = -\frac{1}{2}\int_{\mathbb{R}} f''(u)\,u_x^3\,dx.$$

Hence, integrating over some interval $(0,t)$, we find

$$\int_{\mathbb{R}} u_x(t)^2\,dx + 2\,\epsilon\int_0^t\int_{\mathbb{R}} u_{xx}^2\,dxdt = \int_{\mathbb{R}} u_{0x}^2\,dx - \int_0^t\int_{\mathbb{R}} f''(u)\,u_x^3\,dxdt. \tag{2.9}$$

We multiply (2.9) by δ and add it up with (2.7):

$$\begin{aligned}\delta\int_{\mathbb{R}} u_x(t)^2\,dx &+ 2\,\epsilon\,\delta\int_0^t\int_{\mathbb{R}} u_{xx}^2\,dxdt\\ &= \int_{\mathbb{R}} U(u_0)\,dx - \int_{\mathbb{R}} U(u(t))\,dx + \delta\int_{\mathbb{R}} u_{0x}^2\,dx\\ &\quad - \epsilon\int_0^t\int_{\mathbb{R}} U''(u)\,u_x^2\,dxdt - \delta\int_0^t\int_{\mathbb{R}} f''(u)\,u_x^3\,dxdt - \frac{\delta}{2}\int_0^t\int_{\mathbb{R}} U'''(u)\,u_x^3\,dxdt.\end{aligned}$$

Choosing U given by

$$U(u) = -2\int_0^u\big(f(v) - f(0)\big)\,dv \tag{2.10}$$

the last two terms in the above identity cancel out. Since

$$-c \le \frac{U''(u)}{2} \le c := \|f'\|_\infty \qquad \text{for all } u\in\mathbb{R},$$

thus

$$-c\,u^2 \le U(u) \le c\,u^2 \qquad \text{for all } u\in\mathbb{R},$$

and we finally obtain

$$\begin{aligned}\delta\int_{\mathbb{R}} u_x(t)^2\,dx &+ 2\,\epsilon\,\delta\int_0^t\int_{\mathbb{R}} u_{xx}^2\,dxdt\\ &\le \int_{\mathbb{R}} c\,u_0^2\,dx + \int_{\mathbb{R}} c\,u(t)^2\,dx + \delta\int_{\mathbb{R}} u_{0x}^2\,dx + 2\,\epsilon\int_0^t\int_{\mathbb{R}} c\,u_x^2\,dxdt.\end{aligned}$$

Hence using (2.8)

$$\delta\int_{\mathbb{R}} u_x(t)^2\,dx + 2\,\epsilon\,\delta\int_0^t\int_{\mathbb{R}} u_{xx}^2\,dxdt \le 2\,c\int_{\mathbb{R}} u_0^2\,dx + \delta\int_{\mathbb{R}} u_{0x}^2\,dx,$$

which leads to (2.5c) and (2.5d). The proof of Theorem 2.1 is completed. □

Recall that by Kruzkov' theory, given $u_0 \in L^2(\mathbb{R})$, the Cauchy problem (2.3)–(2.4) admits a unique entropy solution $u \in L^\infty\big(\mathbb{R}_+, L^2(\mathbb{R})\big)$ in the sense of Kruzkov's theory. See [10, 4, 22, 9].

Theorem 2.2. *Assume that, for some constant $C_0 > 0$ independent of ϵ,*

$$\|u_0^\epsilon\|_{L^2(\mathbb{R})} + \sqrt{\delta}\, \|u_{0x}^\epsilon\|_{L^2(\mathbb{R})} \le C_0. \tag{2.11}$$

(1) *As $\epsilon \to 0$ with $\delta = O(\epsilon^2)$ (a subsequence of) the solution u^ϵ of* (2.1)–(2.2) *converges in $L^p_{loc}\big(\mathbb{R}_+, L^q_{loc}(\mathbb{R})\big)$ (for all $1 < p < \infty$ and $1 < q < 2$) toward a weak solution of the problem* (2.3)–(2.4).

(2) *If the stronger condition $\delta = o(\epsilon^2)$ holds, then the limit is the unique entropy solution in the sense of Kruzkov.*

In case (1) a subsequence of u^ϵ (at least) converges strongly, while in case (2) the whole sequence converges strongly. We can conjecture that, in fact, the whole sequence should converge in case (1) as well, but proving it would be very challenging since it requires a uniqueness result of nonclassical entropy solutions. (See also LeFloch [16].)

Proof. We will apply the general convergence framework established by Schonbek [21]. Based on (2.11) and the uniform estimate (2.5a) derived earlier, we can select a subsequence of u^ϵ converging "in the sense" of the Young measures. To apply [21], we only need to control the entropy dissipation measures associated with (2.1). Let U be a smooth function with (at most) linear growth at infinity and, more precisely, such that U' and U'' are uniformly bounded on $\mathbb{R}$. Consider the distribution

$$\Gamma^\epsilon = U(u^\epsilon)_t + F(u^\epsilon)_x,$$

where as usual $F' = U'\, f'$. With obvious notation consider the decomposition

$$\begin{aligned}\Gamma^\epsilon &= \epsilon\, \big(U'(u^\epsilon)\, u_x^\epsilon\big)_x - \epsilon\, U''(u^\epsilon)\, (u_x^\epsilon)^2 + \delta\, \big(U'(u^\epsilon)\, u_{xx}^\epsilon\big)_x - \delta\, U''(u^\epsilon)\, u_x^\epsilon\, u_{xx}^\epsilon \\ &= \Gamma_1^\epsilon + \Gamma_2^\epsilon + \Gamma_3^\epsilon + \Gamma_4^\epsilon.\end{aligned}$$

The estimates below hold for all smooth function $\theta : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ with compact support in (x,t).

Consider first the term Γ_1^ϵ. By the Cauchy–Schwarz inequality, we get

$$\begin{aligned}\left|\int_0^\infty \int_{\mathbb{R}} \Gamma_1^\epsilon\, \theta\, dxdt\right| &= \left|\int_0^\infty \int_{\mathbb{R}} \epsilon\, U'(u^\epsilon)\, u_x^\epsilon\, \theta_x\, dxdt\right| \\ &\le \epsilon\, C\, \|u_x^\epsilon\|_{L^1(\mathbb{R}_+, L^2(\mathbb{R}))}\, \|\theta_x\|_{L^\infty(\mathbb{R}_+, L^2(\mathbb{R}))} \\ &\le C'\, \sqrt{\epsilon} \to 0,\end{aligned} \tag{2.12i}$$

where we used (2.5b). This proves that Γ_1^ϵ converges to zero in the sense of distributions.

Next we simply point out that, by (2.5b) again, the second term Γ_2^ϵ remains uniformly bounded in L^1:

$$\int_0^\infty \int_{\mathbb{R}} |\Gamma_2^\epsilon|\, dxdt \le \frac{1}{2}\, \|u_0^\epsilon\|^2_{L^2(\mathbb{R})}. \tag{2.12ii}$$

To estimate Γ_3 we use (2.5d):

$$\begin{aligned} \left| \int_0^\infty \int_{\mathbb{R}} \Gamma_3^\epsilon \, \theta \, dxdt \right| &= \left| \delta \int_0^\infty \int_{\mathbb{R}} U'(u^\epsilon) \, u_{xx}^\epsilon \, \theta_x \, dxdt \right| \\ &\le \delta \, C \|u_{xx}^\epsilon\|_{L^1(\mathbb{R}_+, L^2(\mathbb{R}))} \, \|\theta_x\|_{L^\infty(\mathbb{R}_+, L^2(\mathbb{R}))} \\ &\le C' \sqrt{\frac{\delta}{\epsilon}} \to 0, \end{aligned} \tag{2.12iii}$$

provided that the mild condition $\delta = o(\epsilon)$ holds. Therefore Γ_3^ϵ tends to zero in the sense of distributions.

Finally, we deal with the last term as follows:

$$\begin{aligned} \left| \int_0^\infty \int_{\mathbb{R}} \Gamma_4^\epsilon \, \theta \, dxdt \right| &= \left| \int_0^\infty \int_{\mathbb{R}} \delta \, U''(u^\epsilon) \, u_x^\epsilon \, u_{xx}^\epsilon \, \theta \, dxdt \right| \\ &\le \delta \, C \|u_{xx}^\epsilon\|_{L^\infty(\mathbb{R}_+, L^2(\mathbb{R}))} \, \|u_x^\epsilon\|_{L^\infty(\mathbb{R}_+, L^2(\mathbb{R}))} \, \|\theta\|_{L^\infty(\mathbb{R} \times \mathbb{R}_+)} \\ &\le C' \, \frac{\sqrt{\delta}}{\epsilon}, \end{aligned} \tag{2.12iv}$$

where we use (2.5b) and (2.5d). The upper bound above tends to zero iff $\delta = o(\epsilon^2)$, in which case we can conclude that Γ_4^ϵ tends to zero in the sense of distributions. Under the weaker assumption $\delta = O(\epsilon^2)$, we see that Γ_4^ϵ is solely bounded in $L^1(\mathbb{R} \times \mathbb{R}_+)$ as is Γ_2^ϵ.

The conclusion (1) of the theorem follows immediately from the uniform bounds (2.12i)–(2.12iv) by applying Schonbek's convergence theory. Her arguments show only that a subsequence of u^ϵ converges and that the limit is a weak solution of (2.3)–(2.4). On the other hand, assuming now the stronger condition $\delta = o(\epsilon^2)$ and restricting attention to *convex functions* U, in view of (2.12i)–(2.12iv) again and the expression of Γ_2^ϵ, we see that the entropy dissipation decomposes in the form

$$\Gamma^\epsilon = \tilde{\Gamma}^\epsilon + \Gamma_2^\epsilon,$$

where $\tilde{\Gamma}^\epsilon \to 0$ in the sense of distributions and Γ^ϵ is a nonpositive bounded measure. This shows that all of the entropy inequalities hold in the limit $\epsilon \to 0$. Thus the limit coincides with the unique entropy solution of the problem. □

3. Multidimensional conservation laws. The estimates and the technique of proof in section 2 do not apply to multidimensional equations, and markedly different arguments are discussed now. Consider the following Cauchy problem:

$$u_t + \sum_{j=1}^d f_j(u)_{x_j} = \epsilon \sum_{j=1}^d u_{x_j x_j} + \delta \sum_{j=1}^d u_{x_j x_j x_j}, \qquad u = u^\epsilon(x,t), \, x \in \mathbb{R}^d, \, t > 0, \tag{3.1}$$

$$u(x,0) = u_0^\epsilon(x), \quad x \in \mathbb{R}^d. \tag{3.2}$$

Provided the initial data u_0^ϵ converge weakly to some limit u_0 (in L^2, say), we will now prove that the solutions of (3.1)–(3.2) converge toward the entropy solution of the associated hyperbolic problem

$$u_t + \sum_{j=1}^d f_j(u)_{x_j} = 0, \qquad u = (x,t),\ x \in \mathbb{R}^d,\ t > 0, \tag{3.3}$$

$$u(x,0) = u_0(x), \quad x \in \mathbb{R}^d. \tag{3.4}$$

Precisely our results are as follows.

Theorem 3.1. *Suppose that the flux-function f is Lipschitz continuous on $\mathbb{R}$ and that the initial data u_0 belong to $L^2(\mathbb{R}^d)$. Then the solution u^ϵ of* (3.1)–(3.2) *satisfies the following a priori estimates:*

$$\|u^\epsilon(t)\|_{L^2(\mathbb{R}^d)} \le \|u_0^\epsilon\|_{L^2(\mathbb{R}^d)}, \qquad t \ge 0, \tag{3.5a}$$

$$\sqrt{2\,\epsilon}\,\|u_x^\epsilon\|_{L^1(\mathbb{R}^+,L^2(\mathbb{R}^d))} \le \|u_0^\epsilon\|_{L^2(\mathbb{R}^d)}, \tag{3.5b}$$

(3.5c)
$$\epsilon\,\|u_{x_j}^\epsilon(t)\|_{L^2(\mathbb{R}^d)} \le \sqrt{d}\,\|f_j'\|_\infty\,\|u_0^\epsilon\|_{L^2(\mathbb{R}^d)} + \epsilon\,\|u_{0x_j}^\epsilon\|_{L^2(\mathbb{R}^d)}, \qquad j = 1,\ldots,d,\ t \ge 0,$$

and

(3.5d)
$$\epsilon^{3/2}\,\|u_{x_jx_k}^\epsilon\|_{L^1(\mathbb{R}^+,L^2(\mathbb{R}^d))} \le \sqrt{d}\,\|f_j'\|_\infty\,\|u_0^\epsilon\|_{L^2(\mathbb{R}^d)} + \epsilon\,\|u_{0x_j}^\epsilon\|_{L^2(\mathbb{R}^d)}, \qquad j,k = 1,\ldots,d.$$

For each $u_0 \in L^2(\mathbb{R}^d)$, the Cauchy problem (3.3)–(3.4) admits a unique entropy solution $u \in L^\infty(\mathbb{R}_+, L^2(\mathbb{R}^d)$ in the sense of Kruzkov. See again [10, 4, 22, 9].

Theorem 3.2. *Assume that, for some constant $C_0 > 0$ independent of ϵ,*

$$\|u_0^\epsilon\|_{L^2(\mathbb{R}^d)} + \epsilon \sum_{j=1}^d \|u_{0x_j}^\epsilon\|_{L^2(\mathbb{R})} \le C_0. \tag{3.6}$$

Then, when $\epsilon \to 0+$ with $\delta = o(\epsilon^2)$, the solution u^ϵ of (3.1)–(3.2) *converges in $L^p_{loc}\big(\mathbb{R}_+, L^q_{loc}(\mathbb{R}^d)\big)$ (for all $1 < p < \infty$ and $1 < q < 2$) toward the unique entropy solution in the sense of Kruzkov of the Cauchy problem* (3.3)–(3.4).

Recall again that the condition $\delta = o(\epsilon^2)$ is sharp since, in the opposite case, nonclassical solutions violating the Kruzkov entropy inequalities could arise in the limit.

Proof of Theorem 3.1. We omit the upper-index ϵ in the following calculation. To derive the L^2 bound (3.5a), we multiply (3.1) by u and get

$$\Big(\frac{|u|^2}{2}\Big)_t + \sum_{j=1}^d F_j(u)_{x_j} = \sum_{j=1}^d \big(\epsilon\,u\,u_{x_j}\big)_{x_j} - \epsilon \sum_{j=1}^d |u_{x_j}|^2 - \frac{\delta}{2} \sum_{j=1}^d \big(|u_{x_j}|^2\big)_{x_j} + \sum_{j=1}^d \big(\delta\,u\,u_{x_jx_j}\big)_{x_j},$$

where $F_j' = u\,f_j'$ is normalized by the condition $F_j(0) = 0$, $j = 1,\ldots,d$. Integrating over space, we get

$$\frac{d}{dt}\int_{\mathbb{R}^d} |u|^2\,dx = -2\,\epsilon \int_{\mathbb{R}^d} \sum_{j=1}^d |u_{x_j}|^2\,dx,$$

and so for all $t \geq 0$

$$\int_{\mathbb{R}^d} |u(t)|^2\, dx + 2\,\epsilon \int_0^t \int_{\mathbb{R}^d} \sum_{j=1}^{d} |u_{x_j}|^2\, dxdt = \int_{\mathbb{R}^d} |u_0|^2\, dx. \tag{3.7}$$

To estimate the gradient of u, for $k = 1, \ldots, d$ we differentiate (3.1) with respect to the variable x_k and then multiply by u_{x_k}. The right-hand side of (3.1) is linear in u thus the calculation for this side is identical to the one we just made, but with u replaced with u_{x_k}. On the other hand, the flux term in the left-hand side is nonlinear and requires a specific argument,

$$\frac{d}{dt} \int_{\mathbb{R}^d} |u_{x_k}|^2\, dx - \sum_{j=1}^{d} \int_{\mathbb{R}^d} 2\, u_{x_k x_j}\, f_j'(u)\, u_{x_k}\, dx = -2\epsilon \sum_{j=1}^{d} \int_{\mathbb{R}^d} |u_{x_j x_k}|^2\, dx,$$

so after integration in time

$$\begin{aligned} &\int_{\mathbb{R}^d} |u_{x_k}(t)|^2\, dx + 2\,\epsilon \sum_{j=1}^{d} \int_0^t \int_{\mathbb{R}^d} |u_{x_j x_k}|^2\, dxdt \\ &\leq \int |u_{0x_k}|^2\, dx + 2\, \|f_k'\|_\infty \sum_{j=1}^{d} \int_0^t \int_{\mathbb{R}^d} |u_{x_j x_k}|\, |u_{x_k}|\, dxdt \\ &\leq \int_{\mathbb{R}^d} |u_{0x_k}|^2\, dx + \frac{\|f_k'\|_\infty^2}{\epsilon}\, d \sum_{j=1}^{d} \int_0^t \int_{\mathbb{R}^d} |u_{x_k}|^2\, dxdt + \epsilon \sum_{j=1}^{d} \int_0^t \int_{\mathbb{R}^d} |u_{x_j x_k}|^2\, dxdt. \end{aligned}$$

Observe that the last term of the right-hand side coincides with the last term of the left-hand side. Therefore, multiplying the above inequality by ϵ^2 and using the entropy dissipation bound in (3.7), we deduce that

$$\begin{aligned} &\int_{\mathbb{R}^d} \epsilon^2\, |u_{x_k}(t)|^2\, dx + \sum_{j=1}^{d} \int_0^t \int_{\mathbb{R}^d} \epsilon^3\, |u_{x_j x_k}|^2\, dxdt \\ &\leq \int_{\mathbb{R}^d} \epsilon^2\, |u_{0x_k}|^2\, dx + \|f_k'\|_\infty^2 \int_0^t \int_{\mathbb{R}^d} d\, \epsilon\, |u_{x_k}|^2\, dxdt \\ &\leq \int_{\mathbb{R}^d} \epsilon^2\, |u_{0x_k}|^2\, dx + d\, \|f_k'\|_\infty^2 \int_{\mathbb{R}^d} |u_0|^2\, dx. \qquad \square \end{aligned} \tag{3.8}$$

Proof of Theorem 3.2. We will rely on the convergence framework proposed by DiPerna [4] for L^∞ solutions and generalized to L^p solutions by Szepessy [22] and Kondo and LeFloch in [9].

Consider a Young measure ν associated with the sequence u^ϵ and based on the uniform L^2 bound (3.5a). (Such Young measures are described in Schonbek [21]). To show that ν is an entropy measure-valued solution, we must check entropy inequalities associated with (3.3), that is,

$$\langle \nu, U \rangle_t + \sum_{j=1}^{d} \langle \nu, F_j \rangle_{x_j} \leq 0, \tag{3.9}$$

where $U : \mathbb{R} \to \mathbb{R}$ is a convex function with (at most) linear growth at infinity and the entropy flux $F_j' = U' f_j'$ is normalized so that $F_j(0) = 0$.

By the definition of the Young measure, we only need to establish that, in the decomposition

$$
\begin{aligned}
\partial_t U(u^\epsilon) + \sum_{j=1}^{d} \partial_j F_j(u^\epsilon) &= \sum_{j=1}^{d} \partial_j \big(\epsilon\, U'(u^\epsilon)\, \partial_j u^\epsilon + \delta(\epsilon)\, U'(u^\epsilon)\, \partial_j^2 u^\epsilon\big) \\
&\quad - \sum_{j=1}^{d} \epsilon\, U''(u^\epsilon)\, |\partial_j u^\epsilon|^2 + \delta(\epsilon)\, U''(u^\epsilon)\, \partial_j u^\epsilon\, \partial_j^2 u^\epsilon \\
&=: \Gamma_1^\epsilon + \Gamma_2^\epsilon + \Gamma_3^\epsilon + \Gamma_4^\epsilon,
\end{aligned}
\tag{3.10}
$$

we have

$$\Gamma_1^\epsilon, \Gamma_2^\epsilon, \Gamma_4^\epsilon \to 0$$

and

$$\Gamma_3^\epsilon \leq 0.$$

These convergence properties were precisely established in the proof of Theorem 2.2, at least for one-dimensional equations. The extension to multidimensional equations is immediate in view of the uniform estimates (3.5a)–(3.5d). A detailed discussion of the initial condition at $t = 0$ (which is based on using suitable entropy inequalities) can be found in Kondo and LeFloch in [9]. This completes the proof that the convergence framework in [9] applies and provides the strong convergence toward the unique entropy solution of (3.3)–(3.4). □

REFERENCES

[1] P. Baiti, P.G. LeFloch, and B. Piccoli, *Uniqueness of classical and nonclassical solutions for nonlinear hyperbolic systems,* J. Differential Equations, 172 (2001), pp. 59–82.

[2] N. Bedjaoui and P.G. LeFloch, *Diffusive-dispersive traveling waves and kinetic relations* I. *Nonconvex hyperbolic conservation laws,* J. Differential Equations, 178 (2002), pp. 574–607.

[3] J. Correia and P.G. LeFloch, *Nonlinear diffusive-dispersive limits for multidimensional conservation laws*, in Advances in Nonlinear P.D.E.'s and Related Areas, G.Q. Chen et al., eds., World Scientific, River Edge, NJ, 1999, pp. 103–123.

[4] R.J. DiPerna, *Measure-valued solutions to conservation laws,* Arch. Ration. Mech. Anal., 88 (1985), pp. 223–270.

[5] B.T. Hayes and P.G. LeFloch, *Nonclassical shocks and kinetic relations: Scalar conservation laws,* Arch. Ration. Mech. Anal., 139 (1997), pp. 1–56.

[6] B.T. Hayes and P.G. LeFloch, *Nonclassical shocks and kinetic relations: Strictly hyperbolic systems,* SIAM J. Math. Anal., 31 (2000), pp. 941–991.

[7] B.T. Hayes and P.G. LeFloch, *Nonclassical shocks and kinetic relations: Finite difference schemes,* SIAM J. Numer. Anal., 35 (1998), pp. 2169–2194.

[8] 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.

[9] C. Kondo and P.G. LeFloch, *Measure-valued solutions and well-posedness of multidimensional conservation laws in a bounded domain,* Portugal. Math., 58 (2001), pp. 171–194.

[10] S.N. Kružkov, *First order quasilinear equations in several independent variables,* Mat. Sb., 81 (1970), pp. 285–355 (in Russian); Math. USSR-Sb., 10 (1970), pp. 217–243 (in English).

[11] P.D. Lax, *The zero dispersion limit, a deterministic analogue of turbulence,* Comm. Pure Appl. Math., 44 (1991), pp. 1047–1056.

[12] P.D. Lax and C.D. Levermore, *The small dispersion limit of the Korteweg-de Vries equation.* I, Comm. Pure Appl. Math., 36 (1983), pp. 253–290.

[13] P.D. Lax and C.D. Levermore, *The small dispersion limit of the Korteweg-de Vries equation.* II, Comm. Pure Appl. Math., 36 (1983), pp. 571–593.

[14] P.D. Lax and C.D. Levermore, *The small dispersion limit of the Korteweg-de Vries equation.* III, Comm. Pure Appl. Math., 36 (1983), pp. 809–829.

[15] P.G. LeFloch, An introduction to nonclassical shocks of systems of conservation laws, Proc. International School on Theory and Numerics for Conservation Laws, Freiburg, Germany, 20–24 Oct. 97, Lectures Notes Comput. Sci. Eng., D. Kröner, M. Ohlberger and C. Rohde, eds., Springer-Verlag, New York, 1999, pp. 28–73.

[16] P.G. LeFloch, *Hyperbolic Systems of Conservation Laws: The Theory of Classical and Nonclassical Shock Waves,* E.T.H. Lecture Notes Series, Birkhäuser, 2002.

[17] P.G. LeFloch and R. Natalini, *Conservation laws with vanishing nonlinear diffusion and dispersion,* Nonlinear Anal., 36 (1999), pp. 213–230.

[18] P.G. LeFloch and C. Rohde, *High-order schemes, entropy inequalities, and nonclassical shocks,* SIAM J. Numer. Anal., 37 (2000), pp. 2023–2060.

[19] P.G. LeFloch and M.D. Thanh, *Nonclassical Riemann solvers and kinetic relations* III. *A non-convex hyperbolic model for van der Waals fluids,* Electron. J. Differential Equations, 72 (2000), 19 pp.

[20] F. Murat, *Compacité par compensation,* Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 5 (1978), pp. 489–507.

[21] M.E. Schonbek, *Convergence of solutions to nonlinear dispersive equations,* Comm. Partial Differential Equations, 7 (1982), pp. 959–1000.

[22] A. Szepessy, *An existence result for scalar conservation laws using measure-valued solutions,* Comm. Partial Differential Equations, 14 (1989), pp. 1329–1350.

[23] L. Tartar, *The compensated compactness method applied to systems of conservation laws*, in Systems of Nonlinear Partial Differential Equations, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. III, J.M. Ball, ed., Reidel, Dordrecht, The Netherlands, 1983, pp. 263–285.

SIAM J. Math. Anal.
Vol. 33, No. 6, pp. 1330–1355

MATHEMATICAL ANALYSIS OF A MODEL FOR THE INITIATION OF ANGIOGENESIS[*]

MARCO A. FONTELOS[†], AVNER FRIEDMAN[‡], AND BEI HU[§]

Abstract. In this paper we consider a nonlinear system of partial differential equations consisting of one parabolic equation and two ordinary differential equations in t. The system arises in a mathematical model of angiogenesis, a process of sprouting of new blood vessels from an existing vascular network. We prove that the system has a unique global solution and study its asymptotic behavior as $t \to \infty$. In particular, we show that stationary solutions are local attractors.

Key words. angiogenesis, tumor growth, degenerate parabolic equations

AMS subject classifications. 35K55, 35K65, 92C15

PII. S0036141001385046

1. Introduction.

The inner lining of blood vessel is made up by a monolayer of flattened and extended cells, called endothelial cells. They are supported by a collagenous network—a network of fibrous proteins. Adhesion molecules, called fibronectins, reside both in the membrane of the blood vessels and in the extracellular matrix that surrounds the vessels.

Angiogenesis is a process of sprouting new blood vessels from a pre-existing vascular network. This process occurs during placental growth, during wound healing, and in tumor growth. In the latter it is initiated by release of chemicals by the tumor, the so called "tumor angiogenic factor," or TAF; see [2]. In a recent paper (cf. [14]) Levine, Sleeman, and Nilsen-Hamilton developed a mathematical model that describes the initiation of angiogenesis in a tumor. The process takes into account the biochemical steps by which chemotactic substances from the tumor combine with the receptors on the endothelial cell wall to produce substances that cause the eventual degradation of the vascular lamina and the migration of endothelial cells to sprout the lining of new capillaries.

The endothelial cell receptors are viewed in this model as the catalyst for transformation of TAF into proteolytic enzyme. The proteolytic enzyme acts as a stimulus for endothelial cell motion via chemotaxis and at the same time as an agent for the degradation of fibronectins. Finally, fibronectin acts as a chemotactic agent for the endothelial cells via a process called haptotaxis. In this process, the endothelial cells tend to move toward low concentrations of fibronectin. In addition, the endothelial cells produce fibronectin and, according to [14], they do it by following a logistic law. The sequence of biochemical processes is described by the law of mass action and the standard Michaelis–Menten kinetics for catalytic reactions, while the cell motion is governed by the continuum equations for reinforced random walks introduced in [15].

[*] Received by the editors February 13, 2001; accepted for publication (in revised form) November 5, 2001; published electronically May 23, 2002.
http://www.siam.org/journals/sima/33-6/38504.html

[†] Departamento de Matemática Aplicada, Escuela Superior de Ciencias Experimentales y Tecnología, Universidad Rey Juan Carlos, C/ Tulipán S/N, 28933-Móstoles, Spain (mafontel@escet.urjc.es).

[‡] Ohio State University, Department of Mathematics, 231 West 18th Avenue, Columbus, OH 43210 (afriedman@math.ohio-state.edu). This author is partially supported by National Science Foundation grant DMS-0098520.

[§] Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556 (Hu.1@nd.edu).

Putting all these facts together and taking the time scale in which some of the biochemical processes are quasi-stationary, one arrives at the following system (see [14] for further details):

$$\eta_t = D_1\eta_{xx} - D_1\left[\eta\left(\ln\tau_1(c,f)\right)_x\right]_x \quad \text{for } 0 \le x \le 1,\ t > 0, \tag{1.1}$$

$$\frac{\partial v}{\partial t} = -\frac{\lambda_1 v\eta}{1+\lambda_2 v} \quad \text{for } 0 \le x \le 1,\ t > 0, \tag{1.2}$$

$$\frac{\partial c}{\partial t} = \frac{\lambda_1 v\eta}{1+\lambda_2 v} \quad \text{for } 0 \le x \le 1,\ t > 0, \tag{1.3}$$

$$\frac{\partial f}{\partial t} = \lambda_3 f(f_M - f)\eta - \frac{\lambda_4 cf}{1+\lambda_5 f} \text{ for } 0 \le x \le 1,\ t > 0, \tag{1.4}$$

where η = endothelial cell density, v = concentration of angiogenic factor, c = concentration of proteolytic enzyme, f = concentration of fibronectin, and D, λ_i, f_M are positive constants. All the functions are also positive. The function $\tau_1(c,f)$ represents a transition probability in the theory of reinforced random walks, and its explicit form for the biological systems under study is unknown. According to experimental observations, however, it must be increasing in c and decreasing in f. Following [13] we shall take

$$\tau_1(c,f) = c^{\gamma_1} f^{-\gamma_2}, \tag{1.5}$$

with γ_i positive constants.

The variable x, after rescaling, is taken along the length of a pre-existing capillary. The boundary condition for η is

$$\eta_x - \eta\left(\ln\tau_1(c,f)\right)_x = 0 \text{ at } x = 0, 1. \tag{1.6}$$

This condition means that there is no flux of cells across the endpoints of the capillary vessel. Finally, initial conditions are prescribed for each of the four unknown functions.

In a subsequent paper (cf. [13]), Levine, Sleeman, and Nilsen-Hamilton incorporate into the model the effects of pericytes (with density $\sigma(x,t)$) and macrophages (with density $m(x,t)$) and of a chemotactic agent (with concentration $u(x,t)$) acting on macrophages. The equations they derive are

$$\sigma_t = D_2\sigma_{xx} - D_2\left[\sigma\left(\ln\tau_2(f)\right)_x\right]_x \quad \text{for } 0 \le x \le 1,\ t > 0, \tag{1.7}$$

$$m_t = D_3 m_{xx} - D_3\left[m\left(\ln\tau_3(c,f)\right)_x\right]_x \quad \text{for } 0 \le x \le 1,\ t > 0 \tag{1.8}$$

for the cells' motion; a new equation for the chemical kinetics of chemotactic agent is

$$\frac{\partial u}{\partial t} = -\frac{\lambda_2 um}{1+\lambda_5 u} \quad \text{for } 0 \le x \le 1,\ t > 0, \tag{1.9}$$

and, finally, the fact that macrophages transform chemotactic agent into TAF leads to the equation

$$\frac{\partial v}{\partial t} = \frac{\lambda_2 um}{1+\lambda_5 u} - \frac{\lambda_1 v\eta}{1+\lambda_2 v} \quad \text{for } 0 \le x \le 1,\ t > 0, \tag{1.10}$$

which replaces (1.2).

We briefly explain the processes underlying the enlarged model; for more details see [13]. Endothelial cells, as mentioned above, reside in the inner part of the membrane of the capillary vessels, and pericyte cells reside on the outer boundary of the membrane. Macrophage cells are located in the extracellular matrix outside the membrane. The tumor secretes angiogenic molecules that go directly to attack the membrane and chemotactic molecules (with concentration u) that interact with macrophages and are converted into angiogenic molecules. The angiogenic molecules (with concentration v) produce proteolytic enzyme (with concentration c) that dilate and break up the membrane, and thereby induces migration of its fibronectin molecules (with concentration f). Pericyte cells tend to move up a fibronectin gradient, while macrophages tend to move up the concentration gradient of the tumor emitted chemotactic factor.

Numerical solutions given in [13] and [14] exhibit the profiles of various cell densities and chemical concentrations during a real-time interval of several days. A large concentration of endothelial cells at one location is associated with the sprouting of a new capillary at that location. Therefore, it is important to know whether the solutions of the above systems lead to aggregation of endothelial cells in some region or, on the contrary, these cells tend to distribute uniformly in the capillary. In the first case, sprouting of new vessels can take place in regions of high aggregation while, in the second case, the model fails to describe this phenomenon.

In this paper we study primarily the model (1.1)–(1.4).

By adding (1.2) and (1.3) we see that $v + c$ is independent of t so that

$$v(x,t) = g(x) - c(x,t),$$

where $g(x)$ is a given function. Following [13], we shall assume that v and f are small enough so that we can approximate $1 + \lambda_2 v$ and $1 + \lambda_5 f$ by 1 and $f_M - f$ by f_M. Rewriting λ_3 instead of $\lambda_3 f_M$, one arrives at the following system:

$$\eta_t = D_1 \eta_{xx} - D_1 \left[\eta \left(\gamma_1 \frac{c_x}{c} - \gamma_2 \frac{f_x}{f} \right) \right]_x, \qquad 0 \le x \le 1,\ t > 0, \tag{1.11}$$

$$\frac{\partial c}{\partial t} = \lambda_1 (g - c)\eta, \qquad 0 \le x \le 1,\ t > 0, \tag{1.12}$$

$$\frac{\partial f}{\partial t} = \lambda_3 f \eta - \lambda_4 c f, \qquad 0 \le x \le 1,\ t > 0, \tag{1.13}$$

together with the no flux boundary condition

$$\eta_x - \eta \left(\gamma_1 \frac{c_x}{c} - \gamma_1 \frac{f_x}{f} \right) = 0 \ \text{ at } x = 0, 1 \tag{1.14}$$

and with positive initial conditions

$$\eta(x,0) = \eta_0(x), \quad c(x,0) = c_0(x), \quad f(x,0) = f_0(x). \tag{1.15}$$

Note that by the definition of $g(x)$ one has $c(x,t) \le g(x)$. Hence, $c(x,t)$ is uniformly bounded and, since $c_t \ge 0$, $\theta(x) = \lim_{t\to\infty} c(x,t)$ exists for any global solution of the system (1.1)–(1.4). This leads one to consider (as in [13]) the subsystem

$$\eta_t = D_1 \eta_{xx} - D_1 \left[\eta \left(\gamma_1 \frac{\theta_x}{\theta} - \gamma_2 \frac{f_x}{f} \right) \right]_x, \qquad 0 \le x \le 1,\ t > 0\ , \tag{1.16}$$

$$\frac{\partial f}{\partial t} = \lambda_3 f \eta - \lambda_4 \theta f, \qquad 0 \le x \le 1,\ t > 0, \tag{1.17}$$

with

$$\eta_x - \eta\left(\gamma_1 \frac{\theta_x}{\theta} - \gamma_2 \frac{f_x}{f}\right) = 0 \text{ at } x = 0, 1,\ t > 0, \tag{1.18}$$

$$\eta(x,0) = \eta_0(x),\ f(x,0) = f_0(x), \tag{1.19}$$

where θ is a *given* positive function. The system (1.16)–(1.19) can also be viewed as a particular case of (1.11)–(1.15) when $c_0(x) \equiv g(x)$.

In this paper we prove that the initial-boundary value problem (1.16)–(1.19) has a unique global classical solution for all $t > 0$ and we establish the phenomenon of aggregation. More specifically, we prove that the nonconstant stationary solution of (1.16)–(1.18) is a local attractor for the time-dependent solution of (1.16)–(1.19). A similar result is proved for the more general system (1.11)–(1.15) provided $g(x) - c_0(x)$ is small enough.

The proofs for (1.11)–(1.15) can actually be modified to include the results for the system (1.16)–(1.19). However, the treatment of this latter system is much simpler and, furthermore, suggests how to approach the more general case of (1.11)–(1.15). For this reason we first give the proofs for the system (1.16)–(1.19) in full details and then consider the general system (1.11)–(1.15), omitting some details.

Our results can also be extended to the system (1.1)–(1.6) with $\tau_1(c, f)$ being a more complicated function like, for instance,

$$\tau_1(c,f) = \left(\frac{\alpha_1 + c}{\alpha_2 + c}\right)^{\gamma_1} \left(\frac{\beta_2 + f}{\beta_1 + f}\right)^{-\gamma_2},$$

where α_i, β_i are positive constants.

The system (1.16)–(1.19) was also considered in [12] for γ_2 both positive and negative. The authors proved that if $\gamma_2 < 0$, then there are solutions (η, f) such that $\eta(x,t)$ becomes infinite in finite time, and if $\gamma_2 > 0$, there are solutions that exist for all time. In this paper we prove that if $\gamma_2 > 0$, then *all* solutions of (1.16)–(1.19) exist for all times. Another proof of this result was given in [16], but our proof is simpler and is the one we are going to adapt to the analysis of system (1.11)–(1.15); see Remark 3.6.

The system (1.16), (1.17) with $\theta = \text{const.}$ is a special case of general chemotaxis equations

$$\frac{\partial p}{\partial t} = \operatorname{div}\left(\nabla p - p\chi(w)\nabla w\right), \tag{1.20}$$

$$\frac{\partial w}{\partial t} = g(p, w). \tag{1.21}$$

These equations describe the rearrangement or movement of living organisms such as cells (e.g., bacteria) with density p under the influence of nondiffusing chemical species with concentration w. The system was introduced by Othmer and Stevens [15] to describe a random walk for p which is reinforced, or biased, by w, whereas at the same time it affects w. Such is the situation, for instance, for common soil bacteria (called myxobacteria) sliding over slime trails. Different choices of $g(p,w)$ and of the chemotactic sensitivity function $\chi(w)$ lead to different conclusions with regard to the profile of p.

For some choices of χ, g global solutions always exist, but for other choices solutions may blow up in finite time. In particular, as mentioned above, if $\gamma_2 < 0$

in (1.16), then there exist solutions of (1.16)–(1.19) that blow up in finite time [12], whereas in our case, where $\gamma_2 > 0$, global solutions always exist [16].

The structure of the paper is as follows: In section 2 we prove that the system (1.16)–(1.19) has a unique classical solution for small time $0 < t < T$. In section 3 we derive a priori bounds which enable us to extend the solution to all $t > 0$. In section 4 we prove that the stationary solution of (1.16)–(1.19) is a local attractor. In section 5 we extend the results of sections 2 and 3 to the general system (1.11)–(1.15) consisting of one diffusion equation and two ODEs together with no flux boundary conditions and initial conditions. Finally, in section 6 we prove that the stationary solutions of (1.11)–(1.15) are also local attractors.

In a recent paper Friedman and Tello [3] proved global existence of solutions of (1.20), (1.21) under general assumptions on g and χ. Their results, however, do not overlap at all with those of the present paper. In particular, the stationary solutions in [3] are all constants, whereas in the present paper the stationary solutions are nonconstant. The methods of the two papers are also entirely different: In the present paper we use integral estimates, whereas in [3] pointwise estimates are derived by comparison arguments.

A two-dimensional model of angiogenesis with one diffusing density and two nondiffusing species was considered in [1]; however, the question of global existence (or blow-up) for that system has not been studied.

We finally mention the Keller–Segal model [9], [10] for chemotaxis: It consists of (1.20) and

$$\alpha \frac{\partial w}{\partial t} = \varepsilon \triangle w + g(p, w) \qquad (\alpha \geq 0,\ \varepsilon > 0)\ . \tag{1.22}$$

In the case $\alpha = 0$, χ constant, and $g(p, w) = p - 1$, solutions may blow up [8] and the precise profile of the blow-up was studied in [5], [6], [7]. On the other hand in the case $\alpha > 0$, $\chi = 1$, and $g(p, w) = -\beta w + \gamma(p - 1)$, global solutions exist for general initial data [4].

In a recent paper Stevens [18] derived the chemotaxis system (1.20), (1.22) by probabilistic analysis. Her paper also reviews (more thoroughly than we do here) the literature of existence and blow-up of solutions of (1.20), (1.22).

2. Local existence for (1.16)–(1.19).

Set

$$\begin{aligned} \tau &= D_1 t, \\ P &= \frac{\lambda_3 \gamma_2}{D_1} \eta, \\ V &= \log\left(\frac{f^{\gamma_2}}{\theta^{\gamma_1}}\right), \\ g(x) &= \frac{\gamma_2 \lambda_4}{D_1} \theta(x). \end{aligned} \tag{2.1}$$

Writing for simplicity t instead of τ, the system (1.16)–(1.19) becomes

$$P_t = \left(P_x + V_x P\right)_x \ \text{ at } 0 \leq x \leq 1,\ t > 0, \tag{2.2}$$

$$V_t = P - g \text{ at } 0 \leq x \leq 1,\ t > 0, \tag{2.3}$$

$$P_x + V_x P = 0 \text{ at } x = 0, 1,\ t > 0, \tag{2.4}$$

$$P(x, 0) = P_0(x),\ V(x, 0) = V_0(x),\ \ 0 < x < 1. \tag{2.5}$$

Set $Q_T = \{0 < x < 1,\ 0 < t < T\}$ and $I = \{0 < x < 1\}$.

We assume that

$$\begin{aligned} &g(x) \in C^{2+\beta}(I), \quad g(x) \geq g_0 > 0, \\ &P_0(x),\ V_0(x) \in C^{2+\beta}(I), \ \text{ and} \\ &P_{0,x} + V_{0,x}P_0 = 0 \ \text{ at } x = 0, 1. \end{aligned} \tag{2.6}$$

It will be convenient later on to recast the system (2.2)–(2.5) in terms of the variables

$$c = e^{-V}, \quad u = Pe^V \ \ (P = cu).$$

Then

$$cu_t = (cu_x)_x + c^2u^2 - g(x)cu \ \text{ for } 0 \leq x \leq 1,\ t > 0, \tag{2.7}$$

$$c_t = -c^2u + gc \text{ for } 0 \leq x \leq 1,\ t > 0, \tag{2.8}$$

$$u_x(0,t) = u_x(1,t) = 0 \text{ for } t > 0, \tag{2.9}$$

$$u(x,0) = u_0(x) = P_0e^{V_0}, \ \ c(x,0) = c_0(x) = e^{-V_0},\ 0 \leq x \leq 1. \tag{2.10}$$

Equation (2.7) can also be written as

$$u_t - u_{xx} - \frac{c_x}{c}u_x = cu^2 - g(x)u\ . \tag{2.11}$$

Theorem 2.1. *If* (2.6) *holds, then there exists a unique solution of* (2.2)–(2.5) *for* $0 < t < T$, *with* $P \in C^{2+\beta, 1+\frac{\beta}{2}}_{x,t}(Q_T)$ *and* $W \in C^{2+\beta, 1+\frac{\beta}{2}}_{x,t}(Q_T)$, *provided* T *is sufficiently small.*

Proof. It suffices to prove the corresponding theorem for the system (2.7)–(2.10). We introduce the space X_T of functions with finite norm

$$\|u\|_{X_T} = |u|_{L^\infty(Q_T)} + |D_xu|_{C^{1+\beta,\frac{\beta}{2}}_{x,t}(Q_T)}$$

and the ball of radius R,

$$B_R = \{u \in X_T,\ \|u\|_T \leq R\}\,,$$

where

$$R = |u_0|_{C^{1+\beta}(I)} + 1.$$

For any $u \in B_R$ we solve (2.8)–(2.10) and find that

$$\|c\|_{X_T} \leq C_1(R).$$

Next we solve

$$\begin{aligned} U_t &= U_{xx} + \frac{c_x}{c}U_x + (cu^2 - g(x)u) \ \text{ in } Q_T, \\ U_x(0,t) &= U_x(1,t) = 0, \quad 0 < t < T, \\ U(x,0) &= u_0, \quad 0 \leq x \leq 1. \end{aligned}$$

Since $c_x \in C^{1+\beta,\frac{\beta}{2}}_{x,t}(Q_T)$, we can use the Schauder estimates (cf. [11]) to conclude that

$$|U|_{C^{2+\beta,1+\frac{\beta}{2}}_{x,t}(Q_T)} \leq C_2(R).$$

But then, by the mean value theorem,

$$|U-u_0|_{L^\infty(Q_T)} + |D_x(U-u_0)|_{C^{1+\beta,\frac{\beta}{2}}_{x,t}(Q_T)} \leq CT^\delta C_2(R)$$

for some $\delta > 0$. It follows that

$$\|U\|_{X_T} < R \tag{2.12}$$

if T is small enough so that $CT^\delta C_2(R) < 1$. Consider the transformation $S : u \to U$. By (2.12), S maps B_R into itself. From the above analysis we also easily see that

$$\|Su_1 - Su_2\|_{X_T} < C_3 T^\delta \|u_1 - u_2\|_{X_T},$$

with a constant C_3 depending on R. Hence S is a contraction in B_R provided T is sufficiently small. It follows that S has a unique fixed point, and together with the corresponding c they form the unique solution of (2.7)–(2.10). □

3. A priori bounds and global existence. Let (u, c) (or (P, V)) be a solution of (2.7)–(2.10) (or (2.2)–(2.5)) for some time interval $0 < t < T$. Note that u, c, P are all positive in $\overline{Q_T}$. We want to derive bounds

$$|u|_{C^{2+\alpha,1+\frac{\alpha}{2}}_{x,t}(Q_T)} + |c|_{C^{2+\alpha,1+\frac{\alpha}{2}}_{x,t}(Q_T)} \leq C(T) \tag{3.1}$$

or, equivalently,

$$|P|_{C^{2+\alpha,1+\frac{\alpha}{2}}_{x,t}(Q_T)} + |V|_{C^{2+\alpha,1+\frac{\alpha}{2}}_{x,t}(Q_T)} \leq C(T), \tag{3.2}$$

where $C(T)$ is a bounded function of T. In this proof we shall use the fact that

$$\int_0^1 P(x,t)dx = \text{const} \equiv \mu. \tag{3.3}$$

This follows by integrating (2.2) over x and using (2.4).

In what follows we shall denote various bounded functions of T by $C(T)$.

Lemma 3.1. *The following estimate holds:*

$$\sup_{0<t<T}\int_I P(x,t)\log P(x,t)dx + \sup_{0<t<T}\int_I (V_x(x,t))^2\,dx + \int\int_{Q_T}\frac{P_x^2}{P}dxdt \leq C(T). \tag{3.4}$$

Proof. Let $h(s) = \int(\log s)ds = s\log s - s$. If we multiply (2.2) by $\log P$ and integrate over Q_t, we get

$$\int_I h(P(x,t))dx + \int\int_{Q_t}\frac{P_x^2}{P}dxdt + \int\int_{Q_t}V_xP_xdxdt = \int_I h(P_0(x))dx\,. \tag{3.5}$$

But, by (2.3),

$$P_x = (V_t)_x + g_x = (V_x)_t + g_x,$$

so that

$$\int\int_{Q_t} V_x P_x dxdt = \left(\frac{1}{2}\int\int_{Q_t}\left(V_x^2\right)_t dxdt + \int\int_{Q_t} V_x g_x dxdt\right)$$

$$= \frac{1}{2}\int_I V_x^2 dx + \int\int_{Q_t} V_x g_x dxdt - \frac{1}{2}\int_I V_{0,x}^2 dx$$

$$\geq \frac{1}{2}\int_I V_x^2 dx - C\int\int_{Q_t} V_x^2 dxdt - C(t+1)\ .$$

Substituting this into (3.5) and using (3.3) we get

$$\int_I P(x,t)\log P(x,t)dx + \int\int_{Q_T}\frac{P_x^2}{P}dxdt + \int_I V_x^2(x,t)dx \leq C(t+1) + C\int\int_{Q_t} V_x^2 dxdt.$$

Since this estimate yields, by Gronwall's inequality,

$$\int\int_{Q_T} V_x^2 dx \leq C(T),$$

(3.4) follows. □

Lemma 3.2. *There exists a constant $C(T)$ such that*

$$\sup_{0<t<T}\int_I P^2(x,t)dx + \int\int_{Q_T} P_x^2 dxdt \leq C(T). \tag{3.6}$$

Proof. If we write (2.2) in the form

$$P_t = (P_x + a(x,t)P)_x\,,$$

then from Lemma 3.1 we have

$$\sup_{0\leq t\leq T}\int_I |a(x,t)|^2\, dx \leq C(T)$$

so that $\int_0^t \left(\int_I \left|a^2\right|^q dx\right)^{\frac{1}{r}} dt \leq C(T)$ with $q = 1, r = 2$. The latter condition allows us to apply Theorem 2.1 in [11, p. 143]. According to that theorem, the solution P satisfies (3.6). We note that Theorem 2.1 in [11] is actually stated for the case of zero Dirichlet data, but the proof for zero Neumann data is the same. □

Lemma 3.3. *There holds*

$$\sup_{Q_T}|u| \leq C(T).$$

Proof. From (3.3) we deduce that

$$|P(x,t)-\mu|^2 \leq \int_0^1 P_x^2(x,t)dx,$$

and using (3.6) we then get

$$\int_0^T \sup_{x\in I} |P(x,t)|^2 \, dt \le C(T). \tag{3.7}$$

Set $\overline{P}(t) = \max_{x\in I} P(x,t)$. Since $cu = P$, (2.11) gives

$$u_t - u_{xx} - \frac{c_x}{c} u_x \le \overline{P}(t)u.$$

The function

$$U(t) = C_0 e^{\int_0^t \overline{P}(t')dt'} \qquad \left(C_0 \ge \max_{x\in I} u(x,0) \right)$$

satisfies

$$U_t = \overline{P}(t)U$$

and is therefore a supersolution. It follows that

$$u(x,t) \le C_0 e^{\int_0^t \overline{P}(t')dt'} \le C(T)$$

by (3.7). □

From (2.8) we see also that c is bounded from above and below, i.e.,

$$c(x,t) \le C(T), \qquad \frac{1}{c(x,t)} \le C(T). \tag{3.8}$$

Equation (2.7) has the form

$$au_t = (cu_x)_x + f,$$

with f bounded by $C(T)$. In the case of $a \equiv 1$ and zero Dirichlet data for u, the following C^α estimate (cf. [11, p. 204, Thm. 10.1]) holds:

$$|u|_{C^{\alpha,\frac{\alpha}{2}}_{x,t}(Q_T)} \le C(T) \ \text{ for some } 0 < \alpha < 1. \tag{3.9}$$

The proof extends to the present case of zero Neumann data. It also extends to the case where a is uniformly bounded above and below by positive constants provided $|a_t| \le C$, since integration by parts of

$$\int\int (u-k)^+ \, au_t dxdt \qquad (k > 0)$$

yields the additional harmless term

$$-\int\int (u-k)^+ \, a_t udxdt.$$

From (2.8) and (3.9) we deduce that

$$|c|_{C^{\alpha,\frac{\alpha}{2}}_{x,t}(Q_T)} \le C(T). \tag{3.10}$$

LEMMA 3.4. *The estimates* (3.1), (3.2) *hold.*

Proof. Set

$$v = cu_x, \; w = c_x.$$

We easily compute that

$$v_t - v_{xx} + \frac{w}{c} v_x = wcu^2 + cuv - g'cu, \tag{3.11}$$

$$w_t = (-2cu + g)w - cv + g'c. \tag{3.12}$$

We write (3.11) in the form

$$bv_t - \left(\frac{1}{c} v_x\right)_x - uv = f,$$

where

$$b = \frac{1}{c} \,, \; f = wu^2 - g'u.$$

By (3.8) and the boundedness of $|c_t|$ we deduce that $|b_t| \leq C(T)$. Since u is bounded and

$$\int_I \left|\frac{c_x}{c}\right|^2 dx \leq \int_I |V_x|^2 \, dx \leq C(T),$$

also

$$\int_I |f(x,t)|^2 \, dx \leq C \int_I |c_x|^2 \, dx \leq C(T) \; .$$

We are now in a position to apply the C^α estimates of Theorem 10.1 of [11, p. 204]. (The condition (7.1) on page 181 holds with $q = 2$.) Here again we use the fact that $|b_t| \leq C(T)$ and make the same remark as we did following (3.9).

Having proved that $v \in C^{\alpha,\frac{\alpha}{2}}_{x,t}(Q_T)$ we can use (3.12) to prove the same for w. It follows that

$$|u_x|_{C^{\alpha,\frac{\alpha}{2}}_{x,t}(Q_T)} \, , \; |c_x|_{C^{\alpha,\frac{\alpha}{2}}_{x,t}(Q_T)} \leq C(T).$$

We can now bootstrap the regularity of u to $C^{2+\alpha,1+\frac{\alpha}{2}}_{x,t}(Q_T)$ by the Schauder estimates applied to (2.7), and next also derive the same regularity for c using (2.8). Finally, we can bootstrap this also to $\alpha = \beta$ (originally α is just some positive number) by again using the Schauder estimates. □

Having established the a priori bounds (3.1), (3.2), we can now extend Theorem 2.1 step-by-step to $0 < t < T$ for any $T > 0$. The size of each time step depends just on the $C^{2+\beta}$ bound on the solution, and thus just on T. We conclude with the following theorem.

THEOREM 3.5. *If* (2.6) *holds, then there exists a unique global solution of* (2.2)–(2.5) *such that P, W are in $C^{2+\beta,1+\frac{\beta}{2}}_{x,t}(Q_T)$ for any $T > 0$.*

Remark 3.6. Theorem 3.5 (in case $g = 0$) was also proved by Rascle [16]. Both proofs of the a priori estimates begin with Lemma 3.2. However, the rest of the proof is much simpler by our method, since we are able to derive quite quickly a uniform bound on u (Lemma 3.4).

4. Convergence to a stationary solution for system (1.16)–(1.19). If we look for stationary solutions of (1.16)–(1.19) we are forced into taking

$$P = P_s(x), \qquad V_x = V_{s,x} = -\frac{P_{s,x}(x)}{P_s(x)}\ .$$

This implies that

$$P_s = \mu + g(x) - \lambda, \text{ with } \lambda \equiv \int_0^1 g(x)dx, \ \ \mu = \int_0^1 P_s(x)dx. \tag{4.1}$$

$$V_{s,x} = -\frac{g'(x)}{\mu + g(x) - \lambda}. \tag{4.2}$$

By (2.3), one must have $V_{s,t} = P_s - g = \mu - \lambda$ and combining this with (4.2), we conclude

$$V_s(x,t) = (\mu - \lambda)t - \log\left(\mu + g(x) - \lambda\right). \tag{4.3}$$

In this section we prove that the stationary solution given by (4.1), (4.2) is a local attractor, i.e., if the initial data of (P, V) are "close" to (P_s, V_s) (in the sense of (4.5) below), then $(P, V) \to (P_s, V_s)$ exponentially fast as $t \to \infty$.

The phenomena whereby nonstationary solutions converge to a nonconstant stationary solution is called aggregation, according to the definition of Othmer–Stevens [15] and Levine–Sleeman [17]. This phenomenon is of particular interest in the context of angiogenesis, for it suggests that new blood vessels will eventually sprout from points where the nonstationary solution achieves its largest values (cf. [12], [13]). The aggregation phenomena is also of interest to general chemotaxis models (cf. [15], [12]).

In what follows we assume that

$$\alpha \equiv \inf_x \left(\mu + g(x) - \lambda\right) > 0. \tag{4.4}$$

Theorem 4.1. *If* (4.4) *holds and*

$$\int_0^1 \left(\left|P_x(x,0) - P_{s,x}(x)\right|^2 + \left|P(x,0) - P_s(x)\right|^2 + \left|V_x(x,0) - V_{s,x}(x)\right|^2\right) dx \le \varepsilon, \tag{4.5}$$

where ε *is positive and sufficiently small, then*

$$\sup_x \left|P(x,t) - P_s(x)\right| \le C\varepsilon e^{-\nu t}\ , \tag{4.6}$$

$$\sup_x \left|V_x(x,t) - V_{s,x}(x)\right| \le C\varepsilon e^{-\nu t}\ , \tag{4.7}$$

and

$$\int_0^1 \left|P_x(x,t) - P_{s,x}(x)\right|^2 dx \le C\varepsilon e^{-\nu t} \quad \textit{for all } t > 0, \tag{4.8}$$

where C *and* ν *are positive constants.*

It will be convenient to state Theorem 4.1 in a different form. To do that, let us introduce the functions

$$\psi(x,t) = \int_0^x \left(P(\xi,t) - \mu - g(\xi) + \lambda\right) d\xi, \tag{4.9}$$

$$w(x,t) = V_x(x,t) - \psi(x,t) + \frac{g'(x)}{\mu + g(x) - \lambda}\ . \tag{4.10}$$

Then, after a straightforward computation, the system (1.16)–(1.19) can be restated as

$$\psi_t - \psi_{xx} - (\mu + g - \lambda)\,\psi - (\mu + g - \lambda)\,w + \frac{g'\psi_x}{(\mu + g - \lambda)} = \psi_x\,(\psi + w) \equiv F, \tag{4.11}$$

$$w_t + (\mu + g - \lambda)\,\psi + (\mu + g - \lambda)\,w - \frac{g'\psi_x}{(\mu + g - \lambda)} = -\psi_x\,(\psi + w) \equiv -F \tag{4.12}$$

for $0 < x < 1$, with boundary conditions

$$\psi(0,t) = \psi(1,t) = 0 \tag{4.13}$$

and initial conditions

$$\psi(x,0) = \psi_0(x), \ \ w(x,0) = w_0(x), \tag{4.14}$$

where $\psi_0 \in C^1\,[0,1]$, $w_0 \in C^0\,[0,1]$.

It is easily seen that Theorem 4.1 can be restated in the following form.

Theorem 4.2. *If* (4.4) *holds and*

$$\int_0^1 \left(\left|D_t\psi_0\right|^2 + \left|D_x\psi_0\right|^2 + w_0^2\right) dx \le \varepsilon, \tag{4.15}$$

where $\varepsilon > 0$ is sufficiently small, then

$$\sup_x |\psi_x(x,t)| \le C\varepsilon e^{-\nu t}\ , \tag{4.16}$$

$$\sup_x \left(|w(x,t)| + |w_t(x,t)|\right) \le C\varepsilon e^{-\nu t}\ , \tag{4.17}$$

and

$$\int_0^1 \left(\left|D_t\psi(x,t)\right|^2 + \left|D_x\psi(x,t)\right|^2\right) dx \le C\varepsilon e^{-\nu t} \ \textit{ for all } t > 0, \tag{4.18}$$

where C and ν are positive constants.

Proof. It will be convenient to rewrite (4.11) in the form

$$\psi_t - (\mu + g - \lambda)\left(\frac{\psi_x}{\mu + g - \lambda}\right)_x - (\mu + g - \lambda)\,(\psi + w) = F. \tag{4.19}$$

Also, by adding (4.11) and (4.12) we get

$$(\psi + w)_t = \psi_{xx}. \tag{4.20}$$

Differentiating (4.19) with respect to t and using (4.20), we arrive at the equation

$$\psi_{tt} - (\mu + g - \lambda)\left(\frac{\psi_{xt}}{\mu + g - \lambda}\right)_x - (\mu + g - \lambda)\,\psi_{xx} = F_t, \tag{4.21}$$

where the left-hand side depends only on ψ. We shall use this equation together with (4.12) to derive our energy estimates on the solution (ψ, w).

If we multiply (4.21) by ψ_t, divide by $(\mu+g-\lambda)$, and integrate with respect to x, we obtain, after integration by parts,

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_0^1 \frac{\psi_t^2}{\mu+g-\lambda}dx + \int_0^1 \frac{\psi_{xt}^2}{\mu+g-\lambda}dx + \frac{1}{2}\frac{d}{dt}\int_0^1 \psi_x^2 dx \\ = \int_0^1 \frac{F_t\psi_t}{\mu+g-\lambda}dx \equiv K_1.\end{aligned} \tag{4.22}$$

Next we multiply (4.21) by ψ, divide by $(\mu+g-\lambda)$, and integrate with respect to x. Using the relation

$$\psi\psi_{tt} = (\psi\psi_t)_t - \psi_t^2,$$

we obtain, after integration by parts,

$$\begin{aligned}\frac{d}{dt}\int_0^1 \frac{\psi\psi_t}{\mu+g-\lambda}dx - \int_0^1 \frac{\psi_t^2}{\mu+g-\lambda}dx + \frac{1}{2}\frac{d}{dt}\int_0^1 \frac{\psi_x^2}{\mu+g-\lambda}dx + \int_0^1 \psi_x^2 dx \\ = \int_0^1 \frac{F_t\psi}{\mu+g-\lambda} \equiv K_2 .\end{aligned} \tag{4.23}$$

Finally, multiplying (4.12) by w and integrating with respect to x, we find that

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_0^1 w^2 dx + \int_0^1 (\mu+g-\lambda)\psi w dx + \int_0^1 (\mu+g-\lambda) w^2 dx \\ - \int_0^1 \frac{g'\psi_x w}{\mu+g-\lambda}dx = -\int_0^1 Fw dx \equiv K_3.\end{aligned} \tag{4.24}$$

We add (4.22) to (4.23) multiplied by ε_1 and (4.24) multiplied by ε_2 $(\varepsilon_1 > 0, \varepsilon_2 > 0)$ to get

$$\frac{1}{2}\frac{dJ}{dt} = I + K, \tag{4.25}$$

where

$$\begin{aligned}J = \int_0^1 \frac{\psi_t^2}{\mu+g-\lambda}dx + \int_0^1 \psi_x^2 dx + 2\varepsilon_1 \int_0^1 \frac{\psi\psi_t}{\mu+g-\lambda}dx \\ + \varepsilon_1 \int_0^1 \frac{\psi_x^2}{\mu+g-\lambda}dx + \varepsilon_2 \int_0^1 w^2 dx,\end{aligned} \tag{4.26}$$

$$\begin{aligned}I = -\int_0^1 \frac{\psi_{xt}^2}{\mu+g-\lambda}dx + \varepsilon_1 \int_0^1 \frac{\psi_t^2}{\mu+g-\lambda}dx - \varepsilon_1 \int_0^1 \psi_x^2 dx \\ -\varepsilon_2 \int_0^1 (\mu+g-\lambda)\psi w dx - \varepsilon_2 \int_0^1 (\mu+g-\lambda) w^2 dx + \varepsilon_2 \int_0^1 \frac{g'\psi_x w}{\mu+g-\lambda}dx,\end{aligned} \tag{4.27}$$

and

$$K = K_1 + \varepsilon_1 K_2 + \varepsilon_2 K_3. \tag{4.28}$$

Set

$$\beta \equiv \sup_x (\mu+g-\lambda), \qquad \gamma \equiv \sup_x |g_x| .$$

Then

$$I \leq -\beta^{-1}\int_0^1 \psi_{xt}^2 dx + \varepsilon_1\alpha^{-1}\int_0^1 \psi_t^2 dx - \varepsilon_1\int_0^1 \psi_x^2 dx$$
$$+ \varepsilon_2\beta\int_0^1 |\psi w|\, dx - \varepsilon_2\alpha\int_0^1 w^2 dx + \varepsilon_2\gamma\alpha^{-1}\int_0^1 |\psi_x w|\, dx.$$

By the Cauchy–Schwarz inequality,

$$\varepsilon_2\beta\int_0^1 |\psi w|\, dx \leq \frac{1}{4}\varepsilon_2\alpha\int_0^1 w^2 dx + \varepsilon_2\frac{\beta^2}{\alpha}\int_0^1 \psi^2 dx,$$

$$\varepsilon_2\gamma\alpha^{-1}\int_0^1 |\psi_x w|\, dx \leq \frac{1}{4}\varepsilon_2\alpha\int_0^1 w^2 dx + \frac{\varepsilon_2\gamma^2}{\alpha^3}\int_0^1 \psi_x^2 dx,$$

and together with Poincare's inequalities

$$\int_0^1 \psi_x^2 dx \geq \pi^2\int_0^1 \psi^2 dx, \tag{4.29}$$

$$\int_0^1 \psi_{xt}^2 dx \geq \pi^2\int_0^1 \psi_t^2 dx, \tag{4.30}$$

we conclude that

$$\begin{aligned} I \leq & -\frac{1}{2}\beta^{-1}\int_0^1 \psi_{xt}^2 dx - \left[\frac{\beta^{-1}\pi^2}{2} - \varepsilon_1\alpha^{-1}\right]\int_0^1 \psi_t^2 dx \\ & - \left[\varepsilon_1 - \varepsilon_2\left(\frac{\gamma^2}{\alpha^3} + \frac{\beta^2}{\pi^2\alpha}\right)\right]\int_0^1 \psi_x^2 dx - \frac{\varepsilon_2\alpha}{2}\int_0^1 w^2 dx. \end{aligned} \tag{4.31}$$

Setting

$$\Phi^2(t) \equiv \int_0^1 \left(\psi_t^2 + \psi_x^2 + w^2\right) dx,$$

we next prove the following lemma.

LEMMA 4.3. *If ε_1 is sufficiently small, then*

$$\delta\Phi^2 \leq J \leq \delta^{-1}\Phi^2, \tag{4.32}$$

where δ is a positive constant.

Proof. Clearly

$$J \geq \beta^{-1}\int_0^1 \psi_t^2 dx + (1+\varepsilon_1\beta^{-1})\int_0^1 \psi_x^2 dx - 2\varepsilon_1\alpha^{-1}\int_0^1 |\psi\psi_t|\, dx + \varepsilon_2\int_0^1 w^2 dx,$$

and

$$2\varepsilon_1\alpha^{-1}\int_0^1 |\psi\psi_t|\, dx \leq \frac{1}{2}\int_0^1 |\psi|^2\, dx + 2\frac{\varepsilon_1^2}{\alpha^2}\int_0^1 |\psi_t|^2\, dx.$$

Again using Poincare's inequality we find that

$$J \geq \left(\frac{1}{\beta} - 2\left(\frac{\varepsilon_1}{\alpha}\right)^2\right)\int_0^1 \psi_t^2 dx + \left(1 - \frac{1}{2\pi^2} + \varepsilon_1\beta^{-1}\right)\int_0^1 \psi_x^2 dx + \varepsilon_2\int_0^1 w^2 dx.$$

Hence the first inequality in (4.32) holds if ε_1 is sufficiently small.

The proof of the second inequality in (4.32) is immediate. $\square$

From (4.31) and Poincare's inequalities we easily find that if ε_1 and $\frac{\varepsilon_2}{\varepsilon_1}$ are sufficiently small, then

$$I \leq -\delta\Phi^2 - \frac{1}{2}\beta^{-1}\int_0^1 \psi_{xt}^2 dx \,, \tag{4.33}$$

where δ is a small positive constant. We can, in fact, take the δ's in (4.32) and (4.33) to be the same.

We now proceed to estimate the K in (4.28). We begin with K_1.

$$\begin{aligned}|K_1| &\leq \frac{1}{\alpha}\int_0^1 |F_t\psi_t|\,dx \leq \frac{1}{\alpha}\sup_x |\psi_t| \int_0^1 |F_t|\,dx \\ &\leq \frac{1}{\alpha}\sup_x |\psi_t| \left[\int_0^1 |\psi_{xt}\,(\psi+w)|\,dx + \int_0^1 |\psi_x\,(\psi+w)_t|\,dx\right].\end{aligned}$$

Since

$$\sup_x |\psi_t| \leq \left(\int_0^1 \psi_{xt}^2 dx\right)^{\frac{1}{2}} ,$$

we obtain

$$\begin{aligned}|K_1| \quad &\leq \quad \frac{1}{\alpha}\,\|\psi_t\|_{L^\infty}\Big\{\,\|\psi_{xt}\|_{L^2}\,\|\psi\|_{L^2} + \|\psi_{xt}\|_{L^2}\,\|w\|_{L^2} \\ &\qquad\qquad + \|\psi_x\|_{L^2}\,\|\psi_t\|_{L^2} + \|\psi_x\|_{L^2}\,\|w_t\|_{L^2}\Big\} \\ &\leq \quad \frac{1}{\alpha}\,\|\psi_{xt}\|_{L^2}^2\,\Phi + \frac{1}{\alpha}\,\|\psi_{xt}\|_{L^2}\,\Phi^2 + \frac{1}{\alpha}\,\|\psi_{xt}\|_{L^2}\,\|\psi_x\|_{L^2}\,\|w_t\|_{L^2}\,.\end{aligned} \tag{4.34}$$

To estimate $\|w_t\|_{L^2}$ we use (4.12):

$$\|w_t\|_{L^2} \leq \beta\,\|w\|_{L^2} + \beta\,\|\psi\|_{L^2} + \frac{\gamma}{\alpha}\,\|\psi_x\|_{L^2} + \|F\|_{L^2} \leq \left(\beta + \frac{\gamma}{\alpha}\right)\Phi + \|F\|_{L^2}\,.$$

Substituting this into (4.34) we get

$$|K_1| \leq \frac{1}{\alpha}\,\|\psi_{xt}\|_{L^2}^2\,\Phi + \left(\frac{1}{\alpha} + \frac{\beta}{\alpha} + \frac{\gamma}{\alpha^2}\right)\|\psi_{xt}\|_{L^2}\,\Phi^2 + \frac{1}{\alpha}\,\|\psi_{xt}\|_{L^2}\,\Phi\,\|F\|_{L^2}\,. \tag{4.35}$$

The estimate of $|K_2|$ is similar; we just replace (one of) the factor $\|\psi_{xt}\|_{L^2}$ by $\|\psi_x\|_{L^2}$ on the right-hand side of (4.35), i.e.,

$$\begin{aligned}|K_2| &\leq \frac{1}{\alpha}\|\psi_x\|_{L^2}\Big\{\,\|\psi_{xt}\|_{L^2}\,\Phi + \left(1+\beta+\frac{\gamma}{\alpha}\right)\Phi^2 + \Phi\|F\|_{L^2}\Big\} \\ &\leq \frac{1}{\alpha}\Big\{\,\|\psi_{xt}\|_{L^2}\,\Phi^2 + \left(1+\beta+\frac{\gamma}{\alpha}\right)\Phi^3 + \Phi^2\|F\|_{L^2}\Big\}.\end{aligned}$$

Finally,

$$K_3 \leq \|F\|_{L^2}\,\|w\|_{L^2} \leq \|F\|_{L^2}\,\Phi.$$

Combining these estimates we find that

$$\begin{aligned}|K| \quad &\le \quad \frac{1}{\alpha}\left\|\psi_{xt}\right\|_{L^2}^2\Phi + \left(\frac{2}{\alpha}+\frac{\beta}{\alpha}+\frac{\gamma}{\alpha^2}\right)\left(\left\|\psi_{xt}\right\|_{L^2}\Phi^2+\varepsilon_1\Phi^3\right)+\frac{\varepsilon_1}{\alpha}\left\|\psi_{xt}\right\|_{L^2}\Phi^2 \\ &\quad + \left(\frac{1}{\alpha}\left\|\psi_{xt}\right\|_{L^2}\Phi+\frac{1}{\alpha}\varepsilon_1\Phi^2+\varepsilon_2\Phi\right)\|F\|_{L^2}.\end{aligned} \tag{4.36}$$

To estimate the L^2 norm of F we observe that

$$\|F\|_{L^2} \le \|\psi\psi_x\|_{L^2}+\|\psi_x w\|_{L^2} \le \sup_x|\psi|\Phi+\sup_x|\psi_x|\,\Phi \le \Phi^2+\sup_x|\psi_x|\,\Phi. \tag{4.37}$$

By Sobolev's inequality and by (4.11),

$$\begin{aligned}\sup_x|\psi_x| \le \|\psi_x\|_{L^2}+\|\psi_{xx}\|_{L^2} &\le \|\psi_x\|_{L^2}+\|\psi_t\|_{L^2}+\beta\|\psi\|_{L^2}+\beta\|w\|_{L^2} \\ &+\frac{\gamma}{\alpha}\|\psi_x\|_{L^2}+\sup_x|\psi_x|\left(\|\psi\|_{L^2}+\|w\|_{L^2}\right).\end{aligned}$$

Hence

$$\sup_x|\psi_x| \le \frac{\|\psi_x\|_{L^2}+\|\psi_t\|_{L^2}+\beta\|\psi\|_{L^2}+\beta\|w\|_{L^2}+\frac{\gamma}{\alpha}\|\psi_x\|_{L^2}}{1-(\|\psi\|_{L^2}+\|w\|_{L^2})} \le \frac{C_1\Phi}{1-\sqrt{2}\Phi} \tag{4.38}$$

(where $C_1 = C_1(\beta,\alpha,\gamma)$) provided $\Phi < 1/\sqrt{2}$. Substituting this into (4.37) we get

$$\|F\|_{L^2} \le \Phi^2 + C_1\frac{\Phi^2}{1-\Phi} \le C_2\Phi^2 \ (C_2 \text{ constant})$$

provided $\Phi < \frac{1}{2\sqrt{2}}$. From this estimate and (4.36) we find that

$$|K| \le \frac{1}{\alpha}\|\psi_{xt}\|_{L^2}^2\Phi + C_3\|\psi_{xt}\|_{L^2}\Phi^2 + C_4\Phi^3, \tag{4.39}$$

where C_3, C_4 are constants.

A substitution of (4.33) and (4.39) into the differential equation (4.25) gives

$$\frac{1}{2}\frac{dJ}{dt} \le -\delta\Phi^2 - \frac{1}{2}\beta^{-1}\int_0^1 \psi_{xt}^2dx + \frac{1}{\alpha}\|\psi_{xt}\|_{L^2}^2\Phi + C_3\|\psi_{xt}\|_{L^2}\Phi^2 + C_4\Phi^3,$$

and since

$$C_3\|\psi_{xt}\|_{L^2}\Phi^2 \le \frac{1}{4}\beta^{-1}\|\psi_{xt}\|_{L^2}^2 + C_3^2\beta\Phi^4,$$

we conclude that, as long as $\Phi(t) < \frac{\alpha}{4\beta}$, one has

$$\frac{1}{2}\frac{dJ}{dt} \le -\delta\Phi^2 + C\Phi^3. \tag{4.40}$$

By Lemma 4.3 we then obtain the inequality

$$\frac{1}{2}\frac{dJ}{dt} \le -\delta_1 J + AJ^{\frac{3}{2}} \qquad (\delta_1, A > 0),$$

so that

$$J(t) \le Ce^{-\nu t} \qquad (\nu \equiv 2\delta_1),$$

provided $AJ^{\frac{1}{2}}(0) < \delta_1$ and $\Phi(0)$ is small enough. Again recalling Lemma 4.3, we conclude that if (4.5) holds, then $\Phi(t)$ will remain small for all $t > 0$, and (4.18) will hold. Finally, (4.16) follows from (4.38), and (4.17) then follows from the differential equation (4.12). □

5. The general system (1.11)–(1.15). Consider the general system (1.11)–(1.15) with given initial conditions. We shall further assume that the initial condition $c(x,0)$ (initial concentration of proteolytic enzyme) satisfies

$$\sup_{0\le x\le 1}\frac{g(x)-c_0(x)}{c_0(x)}<\frac{\gamma_2\lambda_3}{\gamma_1\lambda_1} \tag{5.1}$$

and that

$$\gamma_2\lambda_3<\frac{1}{2}; \tag{5.2}$$

in this connection, we recall from [13] the values of $\lambda_j,\ \gamma_j$:

$$\lambda_1=73,\quad \lambda_3=0.22,\quad \gamma_1=\gamma_2=1.2\ .$$

For these values, $\gamma_2\lambda_3\ =0.\,26$, so that (5.2) is certainly satisfied.

Since $c(x,t)$ is monotone increasing in t, (5.1) implies that

$$M(x,t)\equiv\gamma_2\lambda_3-\gamma_1\lambda_1\frac{g-c}{c}\ge M_0>0, \tag{5.3}$$

where $M_0=\gamma_2\lambda_3-\gamma_1\lambda_1\sup_{0\le x\le 1}\frac{g(x)-c_0(x)}{c_0(x)}$.

We introduce now the notation

$$\tau=D_1t,\qquad P=\frac{\lambda_4}{D_1}\eta\ ,$$
$$W=\log\frac{f^{\gamma_2}}{c^{\gamma_1}}\ .$$

Then

$$W_t=\left(\gamma_2\lambda_3-\gamma_1\lambda_1\frac{g-c}{c}\right)P-\gamma_2\lambda_4c\ .$$

Replacing τ with t, the system for P, W, and c becomes

$$P_t=(P_x+PW_x)_x\quad\text{for } 0\le x\le 1,\ t>0, \tag{5.4}$$
$$c_t=\lambda_1(g-c)P\quad\text{for } 0\le x\le 1,\ t>0, \tag{5.5}$$
$$W_t=\left(\gamma_2\lambda_3-\gamma_1\lambda_1\frac{g-c}{c}\right)P-\gamma_2\lambda_4c\quad\text{for } 0\le x\le 1,\ t>0, \tag{5.6}$$

$$(P_x+PW_x)(x,t)=0\quad\text{at } x=0,1, \tag{5.7}$$

$$P(x,0)=P_0(x),\quad W(x,0)=W_0(x),\quad c(x,0)=c_0(x), \tag{5.8}$$

$$\left(P_{0,x}+P_0W_{0,x}\right)(x)=0\ \text{ at } x=0,1.$$

By (5.3) the coefficient of P in (5.6) is uniformly positive. This assumption is, in fact, crucial; see Remark 5.8.

Theorem 5.1. *Given $P_0(x), W_0(x), c_0(x) \in C^{2+\beta}[0,1]$, there exists a unique global solution of* (5.4)–(5.8) *such that $P(x,t), W(x,t), c(x,t) \in C_{x,t}^{2+\beta,1+\frac{\beta}{2}}(Q_T)$ for any $T < \infty$.*

The rest of the section is devoted to the proof of Theorem 5.1. The proof of local existence is similar to the proof of Theorem 2.1. Thus it remains to establish a priori bounds. More precisely, assuming that a solution exists for $t < T$, T arbitrary, it suffices to establish a priori $C_{x,t}^{2+\beta,1+\frac{\beta}{2}}(Q_T)$ bounds by a constant $C(T)$, where $C(T)$ is a bounded function of T. To derive such bounds we first multiply (5.4) by $\log P$ and integrate in space and time, and obtain, after integration by parts,

$$\int_0^1 P(x,t)\log P(x,t)dx + \int_0^t\int_0^1 \frac{P_x^2(x,t')}{P(x,t')}dxdt' = -\int_0^t\int_0^1 P_x(x,t')W_x(x,t')dxdt' + C.$$

Next we multiply (5.4) by W and integrate in space and time to obtain, after integration by parts,

$$\begin{aligned}\int_0^1 W(x,t)P(x,t)dx - \int_0^t\int_0^1 W_t(x,t')P(x,t')dxdt' + \int_0^t\int_0^1 P(x,t')W_x^2(x,t')dxdt'\\ = -\int_0^t\int_0^1 P_x(x,t')W_x(x,t')dxdt' + C.\end{aligned}$$

Adding the two equations, we get

$$\begin{aligned}\int_0^1 P(x,t)\log P(x,t)dx + \int_0^t\int_0^1 \frac{P_x^2(x,t')}{P(x,t')}dxdt' + \int_0^t\int_0^1 P(x,t')W_x^2(x,t')dxdt'\\ = -\int_0^t\int_0^1 W_t(x,t')P(x,t')dxdt' - 2\int_0^t\int_0^1 P_x(x,t')W_x(x,t')dxdt'\\ -\int_0^1 W(x,t)P(x,t)dxdt + C \equiv J_1 + J_2 + J_3 + C.\end{aligned} \tag{5.9}$$

Our goal now is to establish good-enough estimates for J_1, J_2, and J_3. The estimate for J_3 is immediate: By (5.6) and (5.1)

$$\begin{aligned}J_3 \equiv -\int_0^1 W(x,t)P(x,t)dx &= -\int_0^1 \left(W(x,0) + \int_0^t W_t(x,t')dt'\right)P(x,t)dx\\ &\le -\int_0^1 \left(W(x,0) - \gamma_2\lambda_4\int_0^t c(x,t')dt'\right)P(x,t)dx\\ &\le \left(\sup_x |W(x,0)| + A\gamma_2\lambda_4 t\right)\int_0^1 P(x,t)dx \le C,\end{aligned}$$

where we have used (3.3) (which is also valid also for the present system) and the inequality $c \le g$.

The estimate for J_1 is given in the following lemma.

Lemma 5.2. *There holds that*

$$J_1 \le \gamma_2\lambda_3\left(2\mu^2 t + \mu\int_0^t\int_0^1 \frac{P_x^2(x,t')}{P(x,t')}dxdt'\right). \tag{5.10}$$

Proof. By (5.6),

$$J_1 \leq \gamma_2\lambda_3 \int_0^t \int_0^1 P^2 dx dt'. \tag{5.11}$$

Also

$$\int_0^t \int_0^1 P^2 dx dt' \leq \int_0^t \left(\sup_x P \int_0^1 P dx \right) dt' = \mu \int_0^t \sup_x P(x,t') dt', \tag{5.12}$$

where we have used (3.3). Since $\sqrt{P}$ equals $\sqrt{\mu}$ at some point x, $0 < x < 1$,

$$\sqrt{P} - \mu^{\frac{1}{2}} \leq \frac{1}{2} \left(\int_0^1 \frac{P_x^2}{P} dx \right)^{\frac{1}{2}}$$

so that

$$P \leq \left[\mu^{\frac{1}{2}} + \frac{1}{2} \left(\int_0^1 \frac{P_x^2}{P} dx \right)^{\frac{1}{2}} \right]^2 \leq 2\mu + \frac{1}{2} \int_0^1 \frac{P_x^2}{P} dx.$$

It follows that

$$\mu \int_0^t \sup_x P(x,t') dt' \leq 2\mu^2 t + \frac{1}{2}\mu \int_0^t \int_0^1 \frac{P_x^2}{P} dx dt'.$$

Using this in (5.12) and recalling (5.11), the assertion (5.10) follows. □

In order to estimate J_2, we express P (from (5.5), (5.6)) in the form

$$P = \frac{W_t + \gamma_2\lambda_4 c}{M}. \tag{5.13}$$

Then

$$P_x = \frac{W_{xt} + \gamma_2\lambda_4 c_x}{M} - \frac{PM_x}{M}. \tag{5.14}$$

Inserting (5.14) into J_2 leads to

$$J_2 = -2\int_0^t \int_0^1 \frac{W_{xt}W_x}{M} dx dt' - 2\int_0^t \int_0^1 \frac{\gamma_2\lambda_4 c_x W_x}{M} dx dt' + 2\int_0^t \int_0^1 \frac{PM_x W_x}{M} dx dt'$$

$$= -2\int_0^t \int_0^1 \frac{W_{xt}W_x}{M} dx dt' - 2\int_0^t \int_0^1 \frac{\gamma_2\lambda_4 c_x W_x}{M} dx dt' + 2\gamma_1\lambda_1 \int_0^t \int_0^1 \frac{Pgc_x W_x}{c^2 M} dx dt'$$

$$\equiv J_{21} + J_{22} + J_{23}. \tag{5.15}$$

We estimate each of the J_{2i} in the following lemmas.

LEMMA 5.3. *The term J_{21} satisfies the estimate*

$$J_{21} \leq -2\int_0^1 W_x^2(x,t) dx + C. \tag{5.16}$$

Proof. Notice that

$$J_{21} = -\int_0^t \int_0^1 \frac{(W_x^2)_t}{M} dxdt'.$$

Integration by parts in t yields

$$J_{21} = -\int_0^1 \frac{W_x^2(x,t)}{M} dx - \int_0^t \int_0^1 \frac{M_t W_x^2}{M^2} dxdt' + C$$

$$= -\int_0^1 \frac{W_x^2(x,t)}{M} dx - \gamma_1 \lambda_1 \int_0^t \int_0^1 \frac{g c_t W_x^2}{c^2 M^2} dxdt' + C$$

and the last integral is positive since $c_t > 0$. Then, by (5.5) and the fact that c and M are bounded from above and below ($M < \gamma_2 \lambda_3$), we get

$$J_{21} \leq -\frac{1}{\gamma_2 \lambda_3} \int_0^1 W_x^2(x,t) dx + C.$$

Recalling (5.2), we conclude the inequality (5.16). □

In order to estimate J_{22} and J_{23}, we need to estimate c_x in terms of W_x.

LEMMA 5.4. *The following estimate holds:*

$$|c_x(x,t)| \leq A(t) + B(t)(g - c) \int_0^t |W_x(x,t')|\, dt', \tag{5.17}$$

where $A(t)$ and $B(t)$ are functions bounded in $[0,T]$ for any $T > 0$.

Proof. Eliminating P from (5.5) and (5.6), we deduce the relation

$$\frac{c_t}{\lambda_1(g-c)} = \frac{W_t + \gamma_2 \lambda_4 c}{\gamma_2 \lambda_3 - \gamma_1 \lambda_1 \frac{g-c}{c}}$$

from which we get

$$\frac{\gamma_2 \lambda_3 + \gamma_1 \lambda_1 - \gamma_1 \lambda_1 \frac{g}{c}}{\lambda_1(g-c)} c_t = W_t + \gamma_2 \lambda_4 c\ . \tag{5.18}$$

Let

$$G(c) = \int^c \frac{\gamma_2 \lambda_3 + \gamma_1 \lambda_1 - \gamma_1 \lambda_1 \frac{g}{c}}{\lambda_1(g-c)} dc = -\gamma_1 \ln c - \frac{\gamma_2 \lambda_3}{\lambda_1} \ln{(g-c)}$$

so that

$$\frac{d}{dt} G(c) = \frac{d}{dt} W + \gamma_2 \lambda_4 c.$$

After integration we get

$$G(c) = W + \gamma_2 \lambda_4 \int_0^t c(x,t') dt' + b(x)$$

for some function $b(x)$. Hence

$$-\gamma_1 \frac{c_x}{c} - \frac{\gamma_2 \lambda_3}{\lambda_1} \frac{g_x - c_x}{g - c} = W_x + \gamma_2 \lambda_4 \int_0^t c_x(x,t') dt' + b'(x).$$

This implies that c_x is the solution of the equation

$$c_x+\frac{\gamma_2\lambda_3 g_x}{\gamma_2\lambda_3+\gamma_1\lambda_1-\gamma_1\lambda_1\frac{g}{c}}=\frac{\lambda_1(g-c)}{\gamma_2\lambda_3+\gamma_1\lambda_1-\gamma_1\lambda_1\frac{g}{c}}\left[W_x+b'(x)+\gamma_2\lambda_4\int_0^t c_x(x,t')dt'\right].$$

By iteration we then obtain

$$c_x+\frac{\gamma_2\lambda_3 g_x}{\gamma_2\lambda_3+\gamma_1\lambda_1-\gamma_1\lambda_1\frac{g}{c}}$$

$$=\frac{\lambda_1(g-c)}{\gamma_2\lambda_3+\gamma_1\lambda_1-\gamma_1\lambda_1\frac{g}{c}}\left[W_x+B(x,t)+\gamma_2\lambda_4\int_0^t K(x,t,t')W_x(x,t')dt'\right]$$

for some bounded kernel $K(x,t,t')$ and some bounded function $B(x,t)$. Hence (5.17) follows. □

LEMMA 5.5. *The term J_{22} defined in* (5.15) *satisfies the estimate*

$$J_{22}\le C(t)\left[1+\int_0^1\int_0^t W_x^2(x,t')dxdt'\right]. \tag{5.19}$$

Proof. By the Cauchy–Schwarz inequality and boundedness of M we have

$$J_{22}\le C\left(\int_0^t\int_0^1 c_x^2dxdt'+\int_0^t\int_0^1 W_x^2dxdt'\right).$$

In order to estimate $\int_0^t\int_0^1 c_x^2dxdt'$ we use (5.17) to deduce

$$|c_x(x,t')|^2\le 2A(t')^2+2B(t')^2(g-c)^2\left(\int_0^{t'}|W_x(x,t'')|\,dt''\right)^2$$

$$\le 2A(t')^2+2B(t')^2t'(g-c)^2\left(\int_0^{t'}|W_x(x,t'')|^2\,dt''\right). \tag{5.20}$$

Hence

$$\int_0^t\int_0^1 c_x^2dxdt'\le 2A(t)^2t+2B(t)^2t(g-c)^2\int_0^t\int_0^1|W_x(x,t')|^2\,dxdt',$$

and the inequality (5.19) follows. □

Finally, we estimate the term J_{23}.

LEMMA 5.6. *The following estimate holds:*

$$J_{23}\le\frac{1}{4}\int_0^t\int_0^1 PW_x^2dxdt'+C(t)\left[1+\int_0^1\int_0^t W_x^2(x,t')dxdt'\right]. \tag{5.21}$$

Proof. By the Cauchy–Schwarz inequality and the boundedness of c and M,

$$J_{23}\le\frac{1}{4}\int_0^t\int_0^1 PW_x^2dxdt'+C\int_0^t\int_0^1 Pc_x^2dxdt'$$

for some constant C. In order to estimate $\int_0^t \int_0^1 Pc_x^2 dxdt'$ we make use of (5.20) and (3.3) and deduce

$$\begin{aligned}
&\int_0^1 \int_0^t Pc_x^2 dxdt' \\
&\leq \int_0^1 \int_0^t \left\{ 2A(t')^2 + 2B(t')^2 t'(g-c)^2 \left(\int_0^{t'} |W_x(x,t'')|^2 \, dt'' \right) \right\} P(x,t')dxdt' \\
&\leq \int_0^t 2A(t')^2 dt' + \int_0^1 \left[\int_0^t |W_x(x,t'')|^2 \, dt'' \int_0^t \left[2B(t')^2 t'(g-c)^2 \right] P(x,t')dt' \right] dx \\
&\leq \int_0^t 2A(t')^2 dt' + \left(\int_0^1 \int_0^t |W_x(x,t'')|^2 \, dxdt'' \right) \left(\sup_x \int_0^t \left[2B(t')^2 t'(g-c)^2 \right] P(x,t')dt' \right).
\end{aligned}$$

But

$$\sup_x \int_0^t \left[2B(t')^2 t'(g-c)^2 \right] P(x,t')dt' \leq C \sup_x \int_0^t \lambda_1 (g-c) P(x,t')dt'$$

$$= C \sup_x \int_0^t c_t dt' = C(c(x,t) - c(x,0)) \leq C,$$

where we have used (5.5). Hence

$$J_{23} \leq \frac{1}{4} \int_0^t \int_0^1 PW_x^2 dxdt' + C(t) \left[1 + \int_0^1 \int_0^t W_x^2(x,t')dxdt' \right]$$

for some function $C(t)$ bounded in any interval $[0,T]$. □

We are now in position to prove the following theorem.

THEOREM 5.7. *For any solution to* (5.4)–(5.6) *in* $0 < t < T$ *the following inequalities hold:*

$$\sup_{t \in [0,T]} \int_0^1 W_x^2(x,t)dx \leq C(T), \tag{5.22}$$

$$\int_0^T \sup_x \left(\sqrt{P} - \mu^{\frac{1}{2}} \right)^2 dt \leq C(t). \tag{5.23}$$

Proof. Let us introduce the function of t

$$K[P,W] \equiv \int_0^1 P(x,t) \log P(x,t)dx + \frac{1}{2} \int_0^t \int_0^1 \frac{P_x^2(x,t')}{P(x,t')} dxdt'$$

$$+ \frac{1}{2} \int_0^t \int_0^1 P(x,t')W_x^2(x,t')dxdt' + 2 \int_0^1 W_x^2(x,t')dx.$$

From the previous lemmas it follows that

$$K[P,W] \leq A(t) + B(t) \int_0^t \int_0^1 W_x^2 dxdt'.$$

Hence

$$\int_0^1 W_x^2(x,t)dx \le A(t) + B(t)\int_0^t \int_0^1 W_x^2 dxdt'.$$

An application of Gronwall's inequality leads to the estimate

$$\int_0^t \int_0^1 W_x^2(x,t)dxdt \le C(t),$$

which implies that $K[P,W]$ is bounded by $C(T)$. In particular,

$$\frac{1}{2}\int_0^t \int_0^1 \frac{P_x^2(x,t')}{P(x,t')}dxdt' \le C(t)$$

and

$$\int_0^t \sup_x \left(\sqrt{P} - \mu^{\frac{1}{2}}\right)^2 dt \le \frac{1}{4}\int_0^t \int_0^1 \frac{P_x^2(x,t')}{P(x,t')}dxdt' \le C(t). \qquad \square$$

Proof of Theorem 5.1. Using (5.22), (5.23) we can continue as in section 3 to prove Theorem 5.1. In particular, we derive for $k = e^{-W}$, $u = Pe^W$, and c equations analogous to (2.2) (for u) and (2.3) (for k and c). Then we show the boundedness of the function u (as in Lemma 3.3) and of k, c and derive $C_{x,t}^{\alpha,\frac{\alpha}{2}}$ estimates for these functions. Finally, we consider the system for $v = ku_x$, k_x, c_x analogous to (3.11), (3.12) in order to bootstrap the regularity for u, k, c as in section 3. $\square$

Remark 5.8. If (5.6) is replaced by

$$W_t = -aP - b,$$

where a, b are positive constants, then global solutions may not exist. Indeed, Levine and Sleeman [12] gave examples of solutions that blow up in finite time. The main assumption we made in Theorem 5.1 is the inequality (5.1), which ensures that the coefficient of P in (5.6) is positive.

6. Convergence to a stationary solution in system (1.11)–(1.15).

The system (1.11)–(1.15) possesses the following stationary solution:

$$\begin{aligned} &P_s(x) = \mu + g(x) - \lambda, \text{ with } \lambda \equiv \int_0^1 g(x)dx,\\ &W_{s,x} = -\frac{g'(x)}{\mu + g(x) - \lambda},\\ &c_s(x) = g(x). \end{aligned}$$

Analogously to (4.3), we can compute

$$W_s = W_s(x,t) = (\mu - \lambda)t - \log(\mu + g(x) - \lambda).$$

In this section we prove a result analogous to Theorem 4.1 for system (1.11)–(1.15), namely, the local stability of the stationary solution. By rescaling we can rewrite the system (5.4)–(5.8) (which is equivalent to (1.11)–(1.15)) in the form

$$P_t = (P_x + PW_x)_x \text{ for } 0 \le x \le 1,\ t > 0, \tag{6.1}$$

$$c_t = \kappa_1(g - c)P \text{ for } 0 \le x \le 1,\ t > 0, \tag{6.2}$$

$$W_t = \left(1 - \kappa_2 \frac{g - c}{c}\right)P - c \text{ for } 0 \le x \le 1,\ t > 0, \tag{6.3}$$

together with suitable initial and boundary conditions. It will be convenient to recast (6.3) in the form

$$W_t = P - g - \kappa_2 \frac{g-c}{c} P + (g-c) \quad \text{for } 0 \le x \le 1,\ t > 0, \tag{6.4}$$

where the last two terms on the right-hand side represent a small perturbation.

Introducing functions ψ and w as in (4.9) and (4.10), respectively (with V substituted by W), and $\chi \equiv g - c$, one arrives at the system

$$\psi_t - \psi_{xx} - (\mu + g - \lambda)\,\psi - (\mu + g - \lambda)\,w + \frac{g'}{(\mu + g - \lambda)}\psi_x = \psi_x\,(\psi + w)\,, \tag{6.5}$$

$$\chi_t + \kappa_1\,(\mu + g - \lambda)\,\chi = -\kappa_1 \chi \psi_x, \tag{6.6}$$

$$\begin{aligned} &w_t + (\mu + g - \lambda)\,\psi + (\mu + g - \lambda)\,w - \frac{g'}{(\mu + g - \lambda)}\psi_x \\ &= -\psi_x\,(\psi + w) + \chi_x - \kappa_2 \left[\frac{\chi}{g - \chi}(\psi_x + \mu + g - \lambda)\right]_x . \end{aligned} \tag{6.7}$$

If $\kappa_1 = \kappa_2 = 0$, then the system reduces to the system (4.11), (4.12).

From (6.6) and its x-derivative we obtain the relations

$$\frac{1}{2}\frac{d}{dt}\int_0^1 \chi^2 dx + \int_0^1 (\mu + g - \lambda)\,\chi^2 dx = -\kappa_1 \int \psi_x \chi^2,$$

$$\frac{1}{2}\frac{d}{dt}\int_0^1 \chi_x^2 dx + \int_0^1 (\mu + g - \lambda)\,\chi_x^2 dx = -\kappa_1 \int_0^1 g_x \chi \chi_x dx - \kappa_1 \int (\psi_x \chi)_x\,\chi_x,$$

so that

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\left[\int_0^1 \chi^2 dx + \varepsilon_3 \int_0^1 \chi_x^2 dx\right] \le -\alpha \left[\int_0^1 \chi^2 dx + \varepsilon_3 \int_0^1 \chi_x^2 dx\right] \\ &+ \kappa_1 \left(\gamma \varepsilon_3^{\frac{1}{2}} + \sup|\psi_x| + \|\psi_{xx}\|_{L^2}\right) \left[\int_0^1 \chi^2 dx + \varepsilon_3 \int_0^1 \chi_x^2 dx\right], \end{aligned} \tag{6.8}$$

where γ and α are as in section 4 and ε_3 is a positive number to be chosen small enough.

Combining (6.8) with the estimates analogous to (4.22)–(4.24) in the present context, it is a simple matter to obtain (4.40) with

$$\Phi(t) \equiv \int_0^1 \left(\psi_t^2 + \psi_x^2 + w^2 + \chi^2 + \chi_x^2\right) dx$$

and prove an inequality like (4.32). An estimate for $\|\psi_{xx}\|_{L^2}$ is obtained in the same way we obtained the estimate for $\sup|\psi_x|$ in section 4. This bound is needed in order to control the right-hand side of (6.8). We conclude the following theorem.

Theorem 6.1. *If*

$$\int_0^1 \left(\sum_{i=1}^2 \left| P^{(i)}(x,0) - P_s^{(i)}(x) \right|^2 + \sum_{i=1}^2 \left| \chi^{(i)}(x,0) \right|^2 + \left| W_x(x,0) - W_{s,x}(x) \right|^2 \right) dx \le \varepsilon, \tag{6.9}$$

where ε is positive and sufficiently small, then

$$\sup_x |P(x,t) - P_s(x)| \le C\varepsilon e^{-\nu t}, \tag{6.10}$$

$$\sup_x |W_x(x,t) - W_{s,x}(x)| \le C\varepsilon e^{-\nu t}, \tag{6.11}$$

$$\sup_x |\chi(x,t)| \le C\varepsilon e^{-\nu t}, \tag{6.12}$$

and

$$\int_0^1 |P_x(x,t) - P_{s,x}(x)|^2 \, dx \le C\varepsilon e^{-\nu t} \text{ for all } t > 0, \tag{6.13}$$

where C and ν are positive constants.

Acknowledgments. The first author thanks the hospitality of the Institute for Mathematics and Its Applications at the University of Minnesota and the Department of Mathematics at Ohio State University. The third author also thanks the University of Minnesota for its hospitality during his visit.

REFERENCES

[1] A. R. A. Anderson and M. A. I. Chaplain, *Continuous and discrete mathematical models of tumor-induced angiogenesis,* Bull. Math. Biology, 60 (1998), pp. 857–899.

[2] J. Folkman, *Angiogenesis—retrospect and outlook,* in Angiogenesis: Key Principles—Science—Technology—Medicine, R. Steiner, P. B. Weisz, and R. Langer, eds., Birkhäuser, Basel, 1992.

[3] A. Friedman and J. I. Tello, *Stability of solutions of chemotactic equations in reinforced random walks*, J. Math. Anal. Appl., to appear.

[4] H. Gajewsky and K. Zacharias, *Global behavior of a reaction-diffusion system modeling chemotaxis,* Math. Nachr., 195 (1998), pp. 177–194.

[5] M. A. Herrero and J. J. L. Velázquez, *Chemotactic collapse for the Kelle-Segel model,* J. Math. Biol., 35 (1996), pp. 177–194.

[6] M. A. Herrero and J. J. L. Velázquez, *Singularity patterns in a chemotactic model,* Math. Ann., 206 (1996), pp. 583–623.

[7] M. A. Herrero and J. J. L. Velázquez, *A blow-up mechanism for a chemotaxis model,* Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 24 (1997), pp. 633–683.

[8] W. Jäger and S. Luckhaus, *On explosions of solutions to systems of partial differential equations modeling chemotaxis,* Trans. Amer. Math. Soc., 329 (1992), pp. 819–824.

[9] E. F. Keller and L. A. Segal, *Initiation of slime mold aggregation viewed as an instability,* J. Theoret. Biol., 26 (1970), pp. 399–415.

[10] E. F. Keller and L. A. Segal, *A model for chemotaxis,* J. Theoret. Biol., 30 (1971), pp. 225–234.

[11] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type,* Transl. Math. Monogr. 23, American Mathematical Society, Providence, RI, 1968.

[12] H. A. Levine and B. D. Sleeman, *A system of reaction diffusion equations arising in the theory of reinforced random walks,* SIAM J. Appl. Math., 57 (1997), pp. 683–730.

[13] H. A. Levine, B. D. Sleeman, and M. Nilsen-Hamilton, *A mathematical model for the roles of pericytes and macrophages in the initiation of angiogenesis* I. *The role of protease inhibitors in preventing angiogenesis,* Math. Biosci., 168 (2000), pp. 77–115.

[14] H. A. Levine, B. D. Sleeman, and M. Nilsen-Hamilton, *Mathematical modeling of the onset of capillary formation initiating angiogenesis,* J. Math. Biol., 42 (2001), pp. 195–238.

[15] H. G. Othmer and A. Stevens, *Aggregation, blowup, and collapse: The ABC's of taxis in reinforced random walks,* SIAM J. Appl. Math., 57 (1997), pp. 1044–1081.

[16] M. Rascle, *Sur une équation intégro-differentielle non linéaire issue de la biologie,* J. Differential Equations, 32 (1979), pp. 420–453.

[17] B. D. Sleeman and H. A. Levine, *Partial differential equations of chemotaxis and angiogenesis,* Math. Methods Appl. Sci., 24 (2001), pp. 405–426.

[18] A. Stevens, *The derivation of chemotaxis equations as limit dynamics of moderately interacting stochastic many-particle systems,* SIAM J. Appl. Math., 61 (2000), pp. 183–212.

SIAM J. MATH. ANAL.
Vol. 33, No. 6, pp. 1356–1378

LINEAR INSTABILITY OF SOLITARY WAVE SOLUTIONS OF THE KAWAHARA EQUATION AND ITS GENERALIZATIONS[*]

THOMAS J. BRIDGES[†] AND GIANNE DERKS[†]

Abstract. The linear stability problem for solitary wave states of the Kawahara—or fifth-order KdV-type—equation and its generalizations is considered. A new formulation of the stability problem in terms of the symplectic Evans matrix is presented. The formulation is based on a multisymplectification of the Kawahara equation, and leads to a new characterization of the basic solitary wave, including changes in the state at infinity represented by embedding the solitary wave in a multiparameter family. The theory is used to give a rigorous geometric sufficient condition for instability. The theory is abstract and applies to a wide range of solitary wave states. For example, the theory is applied to the families of solitary waves found by Kichenassamy–Olver and Levandosky.

Key words. solitary waves, Evans function, multisymplectic structures

AMS subject classifications. 35, 53, 76

PII. S0036141099361494

1. Introduction.

The Kawahara equation—or *fifth-order KdV-type equation*—is a model equation for plasma waves, capillary-gravity water waves, and other dispersive phenomena when the cubic KdV-type dispersion is weak. Such equations can be written in the general form

$$2\frac{\partial u}{\partial t}+\alpha\frac{\partial^3 u}{\partial x^3}+\beta\frac{\partial^5 u}{\partial x^5}=\frac{\partial}{\partial x}f(u,u_x,u_{xx})\,, \tag{1.1}$$

for the scalar-valued function $u(x,t)$, where α and β are real parameters with $\beta\neq 0$ and $f(u,u_x,u_{xx})$ is some smooth function. In many applications the phenomena which lead to the model equation (1.1) have a Hamiltonian structure. Therefore it is natural to require that f be a variational derivative, in which case (1.1) is a Hamiltonian system

$$\frac{\partial u}{\partial t}=\mathcal{J}\frac{\delta\mathcal{H}}{\delta u}\,,\quad\text{with}\quad \mathcal{J}=-\frac{1}{2}\frac{\partial}{\partial x}\,, \tag{1.2}$$

and

$$\mathcal{H}(u)=\int_{\mathbb{R}}\left(\tfrac{1}{2}\beta u_{xx}^2-\tfrac{1}{2}\alpha u_x^2+h(u,u_x,u_{xx})\right)\,dx, \tag{1.3}$$

where the variational derivative of the functional associated with $h(u,u_x,u_{xx})$ yields f. Precise forms for f and h will be given in section 2.

The form of (1.1) which occurs most often in applications is with $f(u,u_x,u_{xx})=a\,u^2$, where a is a nonzero constant. The first appearance of this equation known to the authors is in the Japanese literature: Kawahara [24] points out that Kakutani and Ono suggested the inclusion of a fifth-order term to KdV to model magneto-acoustic waves in 1969, and Hasimoto first showed in 1970 that a fifth-order term was

[*] Received by the editors September 20, 1999; accepted for publication (in revised form) January 13, 2002; published electronically May 23, 2002.
http://www.siam.org/journals/sima/33-6/36149.html

[†] Department of Mathematics and Statistics, University of Surrey, Guildford, Surrey, GU2 7XH, UK (T.Bridges@surrey.ac.uk, G.Derks@surrey.ac.uk).

necessary to model capillary-gravity waves for Bond number near one third. Kawahara [24] appears to have been the first to write down the complete equation (1.1) with $f(u, u_x, u_{xx}) = -3\,u^2$ (see equation (1) in [24]), begin a systematic study, observe that the solitary wave states could have oscillatory tails, and compute examples of such waves numerically. A more general nonlinearity was derived for water waves by Olver [28], using Hamiltonian perturbation theory, with further generalization given by Craig and Groves [13]. Kichenassamy and Olver [25] suggested taking the most reasonable general form for f—including nongradient forms—and then deducing under what conditions explicit solitary wave solutions exist, and Levandosky [26] proposed an interesting class of homogeneous nonlinearities. All of the above proposed nonlinearities can be characterized in the form (1.1), and when f is variational the system has the Hamiltonian formulation (1.2).

The system (1.1) has many classes of solutions, but a class of great interest is solitary wave states that are biasymptotic to a constant state at infinity. Depending on the form of the nonlinearity, the system can also have travelling fronts (a simple example is given on page 452 of [15]) as well as solitary waves biasymptotic to invariant manifolds more complex than the lines to be considered here (cf. section 2 and the comments in section 8 of [7]). However, for definiteness, we will restrict attention here to classes of solitary waves which decay exponentially to a constant (in general nonzero) at infinity. Such solitary waves travelling at speed c (i.e., $u(x,t) = \hat{u}(x-ct)$) satisfy the fourth-order ordinary differential equation

$$\beta\hat{u}_{xxxx} + \alpha\hat{u}_{xx} - 2c\hat{u} - f(\hat{u}, \hat{u}_x, \hat{u}_{xx}) = A\,, \tag{1.4}$$

where A is a constant of integration. When f is a gradient operator it is easily shown that (1.4) is the Euler–Lagrange equation associated with a Lagrange functional, and the Legendre transform of this functional results in a Hamiltonian formulation for (1.4),

$$U_x = J\,\nabla H(U)\,, \quad U \in \mathbb{R}^4\,, \tag{1.5}$$

where J is a standard unit symplectic operator on $\mathbb{R}^4$, and an expression for $H(U)$ is easily deduced but is not needed here. (A more general derivation of such finite-dimensional Hamiltonian system will be a consequence of multisymplectic formulation in section 2.)

Note that the Hamiltonian structure of this ODE with x considered as an evolution direction is distinct and dramatically different from the infinite-dimensional Hamiltonian structure associated with the time direction (1.2). The interplay between these two distinct structures will play an important role in what follows.

The Hamiltonian structure (1.5) of the reduced system (1.4) has been the basis of many of the methods for finding solitary wave states. A review article on the known classes of solitary wave states with an exhaustive list of references is given by Champneys [11]. Also of interest in this paper are the class of solitary waves found by Kichenassamy and Olver [25] and the recent results of Levandosky [26]. In [25], a classification of admissible expressions for f which lead to explicit sech^2 solitary wave states is given. In [26], an energy-momentum argument is used to prove the existence of a class of solitary waves associated with a homogeneous nonlinearity, and in Groves [19], the mountain-pass lemma is used to prove the existence of solitary waves including multibump solitary waves for a class of homogeneous nonlinearities.

Given the existence of such a large range of solitary wave states for (1.4), a natural question is to determine whether they are stable or unstable. The most successful approaches for studying the stability of KdV and generalized KdV (i.e., (1.1) with $\beta = 0$) have been the energy-momentum method for establishing nonlinear stability and instability (Benjamin [2], Bona [3]) and the connection between the derivative of the momentum with respect to the wave speed and stability (Bona, Souganidis, and Strauss [4], Pego and Weinstein [29]). These energy-momentum based methods have been extended to apply to the stability of solitary waves for the fifth-order KdV by several authors.

The momentum for (1.1) can be expressed as

$$\mathcal{I}(u) = \int_{\mathbb{R}} u^2 \, dx \,,$$

and therefore solitary wave states can be characterized as solutions of $\delta\mathcal{H} = c\delta\mathcal{I}$, i.e., as critical points of the Hamiltonian restricted to level sets of the momentum with c as a Lagrange multiplier. The nondegeneracy condition for this constrained variational principle is

$$\frac{d}{dc}\mathcal{I}(\hat{u}) \neq 0 \,,$$

where $\hat{u}(x, c)$ is the family of solitary waves parametrized by c. Rigorous Lyapunov stability can be obtained by proving that $\hat{u}$ is indeed a minimizer for this variational principle. This approach has been very successful for KdV-type equations but is very difficult to generalize to higher-order equations and systems of evolutionary PDEs. However, for some range of parameters and forms for f, Lyapunov-type energy-momentum arguments have been successfully applied to (1.1). The first results of this type are given by Ill'ichev and Semenov [21] for the waves of depression when $\alpha < 0$ which travel at speed $-c$. Karpman [23] shows that when $\alpha = +1$, $\beta < 0$, and $f = -\frac{u^{p+1}}{p+1}$ the energy-momentum argument and the sign of $\frac{dI}{dc}$ precisely determine stability and instability. However, this theory relies on a hypothesis that a certain linear operator has exactly one negative eigenvalue which is difficult to verify in general. Karpman's theory is applied by Dey, Khare, and Kumar [15] to a class of exact solutions, but it appears that this class of solutions is explicit only for isolated values of c (see further comments on this at the end of section 3).

Using the energy-momentum method and a compensated compactness argument, Levandosky [26] proves the existence of solitary waves for a homogeneous nonlinearity and obtains rigorous stability and instability results using an energy-momentum argument and the sign of $\frac{dI}{dc}$ for a restricted range of parameter space. Recently, Dias and Kuznetsov [16] have obtained rigorous lower bounds on the Hamiltonian function for (1.1) when $f = -u^2$ for the solitary waves with oscillatory tails known to exist near the minimum of the dispersion curve, suggesting that at least one of these families of waves is stable.

For general PDEs (not necessarily Hamiltonian), the most successful approach for the analysis of the linear stability problem is based on the Evans function. The Evans function is a complex analytic function of the spectral parameter, and under suitable hypotheses the zeros of the Evans function correspond to eigenvalues (cf. Evans [17], Alexander, Gardner, and Jones [1]). In Bridges and Derks [7], [9], [10] the concept of the *symplectic Evans function* and the *symplectic Evans matrix* were introduced for Hamiltonian evolution equations. This theory, which will be used as a basis for analyzing the linear stability problem for (1.1), will be summarized in sections 2–3. Essentially, the Hamiltonian PDE is reformulated as a Hamiltonian system on a multi-symplectic structure, where a distinct symplectic structure is assigned for the time and space directions (cf. Bridges [5], [6]). This decomposition allows for a geometric analysis of each step of the existence and linear stability analysis and can be used to deduce an explicit geometric condition for linear instability.

The purpose of this paper is fourfold: first, in section 2, we show that the natural geometric structure of (1.1) is not as a Hamiltonian system as in (1.2) but as a Hamiltonian system on a multisymplectic structure. The problem with (1.2) is that it does not encode any information about the spatial Hamiltonian structure (1.5) that arises when looking for solitary waves and in the linearization about a solitary wave. This geometry should be useful in other analyses of (1.1). Second, in section 3, we show that—with α, β, and f fixed—all existing solitary wave solutions come in three-parameter families, and these families are a natural consequence of the geometric structure. One of the parameters is c, the wave speed, and the other two are related to a space-time drift along an affine group orbit, and when nonzero, they lead to a nontrivial constant state at infinity, and they encode information about the linear stability problem (cf. section 5). We have not found any nontrivial effect on stability of the additional parameters, but we consider only a few examples here. (Examples where nontriviality of the state at infinity affects stability can be found in [7], [9].) These additional parameters are an intrinsic part of the geometry of the PDE. Third, in sections 4–6, we formulate the symplectic Evans matrix for this system. This matrix is of interest because zeros of the determinant of the symplectic Evans matrix in the right-half complex plane correspond to unstable eigenvalues. Fourth, in section 8, we present a rigorous geometric condition for instability for a class of solitary wave states of (1.1) based on the theory in [9], and then, in sections 9–10, this geometric instability criterion is applied to two examples of families of solitary wave states in the literature.

2. Multisymplectic structure of the Kawahara equation.

The starting point for the analysis is the Kawahara equation and its generalizations (1.1), where f is any function which can be written as the variational derivative of a functional

$$\frac{1}{2}\int_{\mathbb{R}}[h(u,u_x,u_{xx})]\,dx.$$

A straightforward calculation shows that this implies that h has to be of the form

$$h(q,r,s)=F(q,r)+sE(q,r)$$

and therefore

$$f(q,r,s)=F_q(q,r)-rF_{qr}(q,r)-sF_{rr}(q,r)+2sE_q(q,r)+srE_{rq}(q,r)+r^2E_{qq}(q,r)\,. \tag{2.1}$$

This expression for f includes all the nonlinearities in variational form for (1.1) encountered in the literature including [12], [23], [24], [25], [26], and [28].

Levandosky [26] considers (1.1) with the restriction that $E=0$ and $F(q,r)$ is three-times continuously differentiable and homogeneous of degree $p+1$ for some $p>1$; that is,

$$F(\lambda q,\lambda r)=\lambda^{p+1}F(q,r)$$

for all $\lambda\geq 0$ and $(q,r)\in\mathbb{R}^2$.

Kichenasammy and Olver [25] consider the existence of solitary waves for a generalized Kawahara equation, where they assume the existence of a smooth function

$g(u)$ and constants A, B such that

$$f(q,r,s) = Ar^2 + Bsq + g'(q)\,.$$

They show that a necessary and sufficient condition for the existence of sech^2-type solitary wave solutions is that $g'(q)$ be a cubic polynomial. On the other hand, this function f has a variational structure if and only if $2A = B$, and in this case, the function f can be derived from $h(q,r,s)$ by taking $E = 0$ and $F(q,r) = -Aqr^2 + g(q)$.

To reformulate (1.1) with the variational condition (2.1) on f, as a Hamiltonian system on a multisymplectic structure, we introduce the potential function $q_1(x,t)$, defined by $u = \frac{\partial q_1}{\partial x}$. Then with

$$\begin{array}{lllllll} & q_2 & = & u = \frac{\partial q_1}{\partial x}\,, & p_1 & = & \frac{\partial q_1}{\partial t} - \frac{\partial p_2}{\partial x} - \frac{\partial}{\partial q_2}F - \frac{1}{\beta}E\frac{\partial}{\partial q_2}E - \frac{p_3}{\beta}\frac{\partial}{\partial q_2}E\,,\\ (2.2) & q_3 & = & u_x = \frac{\partial q_2}{\partial x}\,, & p_2 & = & -\alpha q_3 - \frac{\partial p_3}{\partial x} - \frac{\partial}{\partial q_3}F - \frac{1}{\beta}E\frac{\partial}{\partial q_3}E - \frac{p_3}{\beta}\frac{\partial}{\partial q_3}E\,,\\ & & & & p_3 & = & \beta\frac{\partial q_3}{\partial x} - E\,, \end{array}$$

(1.1) reduces to

$$\frac{\partial q_2}{\partial t} + \frac{\partial p_1}{\partial x} = 0\,. \tag{2.3}$$

Combining (2.2) and (2.3), the PDE (1.1) can be written in the form

$$\mathbf{M}Z_t + \mathbf{K}Z_x = \nabla S(Z)\,, \quad Z \in \mathbb{R}^6\,, \tag{2.4}$$

where

$$Z = \begin{pmatrix} q_1\\ q_2\\ q_3\\ p_1\\ p_2\\ p_3 \end{pmatrix}, \quad \mathbf{M} = \begin{bmatrix} 0 & -1 & 0 & 0 & 0 & 0\\ 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} 0 & 0 & 0 & -1 & 0 & 0\\ 0 & 0 & 0 & 0 & -1 & 0\\ 0 & 0 & 0 & 0 & 0 & -1\\ 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}, \tag{2.5}$$

and

$$S(Z) = \tfrac{1}{2}\alpha q_3^2 + \frac{1}{2\beta}p_3^2 + p_1q_2 + p_2q_3 + F(q_2,q_3) + \frac{1}{2\beta}\left(2p_3 + E(q_2,q_3)\right)E(q_2,q_3). \tag{2.6}$$

The skew-symmetric operators $\mathbf{M}$ and $\mathbf{K}$ define the two-forms

$$\begin{aligned} \omega &= \mathbf{d}q_2 \wedge \mathbf{d}q_1,\\ \kappa &= \mathbf{d}p_1 \wedge \mathbf{d}q_1 + \mathbf{d}p_2 \wedge \mathbf{d}q_2 + \mathbf{d}p_3 \wedge \mathbf{d}q_3, \end{aligned} \tag{2.7}$$

with

$$\omega(\xi_1,\xi_2) = \langle \mathbf{M}\xi_1, \xi_2\rangle \quad \text{and} \quad \kappa(\xi_1,\xi_2) = \langle \mathbf{K}\xi_1,\xi_2\rangle\,, \tag{2.8}$$

where $\langle\cdot,\cdot\rangle$ is a standard inner product on $\mathbb{R}^6$. The induced norm is denoted by $\|\cdot\|$. The symplectic form κ is a canonical symplectic structure on $\mathbb{R}^6$ associated with the x-direction, and ω is a rank 2 symplectic structure associated with the t-direction.

There are two symmetries of (2.4) which will be of interest in what follows: the spatial translation invariance (in x) of the system (i.e., the fact that ω, κ, and $S(Z)$ do not depend explicitly on x), and the affine symmetry associated with the fact that q_1 is a potential function.

Let G be the one-parameter affine group associated with this potential symmetry with action

$$\mathcal{G}_\theta Z = Z + \theta V \quad \text{for all } \theta \in \mathbb{R}, \quad \text{where} \quad V = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{2.9}$$

Then the system (2.4) is G-equivariant; that is, $S(Z)$ and the two-forms ω and κ are G-invariant.

A solitary wave state of (2.4) will be composed of two parts. The first part is the shape of the solitary wave which connects the asymptotic states at plus and minus infinity, which will be characterized as a heteroclinic orbit in the phase space $\mathbb{R}^6$. The second part is the state at infinity which will be characterized as an invariant manifold of relative equilibria associated with the group G.

To define the invariant manifold at infinity we use the theory in section 2 of [9]. First note that, with $P(Z) = q_2$ and $Q(Z) = p_1$, the functions $P(Z)$ and $Q(Z)$ satisfy

$$\mathbf{M}V = \nabla P(Z) \quad \text{and} \quad \mathbf{K}V = \nabla Q(Z); \tag{2.10}$$

that is, $P(Z)$, respectively, $Q(Z)$, are the functions which generate the ω-, respectively, κ-, symplectic flow of the group G. The state at infinity is taken to be of the form

$$Z(x,t) = \mathcal{G}_{\theta(x,t)} Z_0(a,b) \quad \text{with} \quad \theta(x,t) = at + bx + \theta_0. \tag{2.11}$$

The point $Z_0 \in \mathbb{R}^6$ and the parameters a and b are defined by the constrained variational problem: find critical points of $S(Z)$ restricted to level sets of the functions P and Q, or

$$\nabla S(Z_0) = a \nabla P(Z_0) + b \nabla Q(Z_0)\,, \quad \text{with} \quad P(Z_0) = \mathcal{P},\ Q(Z_0) = \mathcal{Q}. \tag{2.12}$$

This equation is easily solved to find

$$Z_0 = \begin{pmatrix} q_1^0 \\ b \\ 0 \\ a - F_q(b,0) \\ -F_r(b,0) \\ -E(b,0) \end{pmatrix} \quad \text{with} \quad \begin{aligned} P(Z_0) &= q_2^0 = b = \mathcal{P}, \\ Q(Z_0) &= p_1^0 = a - F_q(b,0) = \mathcal{Q}, \end{aligned} \tag{2.13}$$

and q_1^0 is arbitrary (due to the group action). This state is nondegenerate as a solution of the constrained variational problem since $\frac{\partial(P,Q)}{\partial(a,b)} = -1 \neq 0$.

Let $Z_0 \in \mathbb{R}^6$ be any nondegenerate solution of (2.12) with $q_1^0 = 0$. Then the invariant manifold at infinity is defined to be the following line in $\mathbb{R}^6$ through Z_0:

$$\mathcal{M}_\infty = \{\, Z_0 + \theta V \ : \ \theta \in \mathbb{R}\,\}. \tag{2.14}$$

The solitary wave state will be taken to be biasymptotic to this manifold and of the form

$$Z(x,t) = \mathcal{G}_{\theta(x,t)}[Z_0^- + \mathcal{T}_{\tau(t)} \widehat{Z}(x,a,b,c)], \tag{2.15}$$

where $\mathcal{G}_{\theta(x,t)}$ is as defined in (2.11); Z_0^- is any nondegenerate solution of (2.12) with $q_1^0 = 0$ (the "−" superscript indicates that this is the asymptotic point on $\mathcal{M}_\infty$ as $x \to -\infty$);

$$\mathcal{T}_\tau \widehat{Z}(x,a,b,c) \overset{\text{def}}{=} \widehat{Z}(x-\tau,a,b,c),$$

and $\tau(t) = ct + \tau_o$. The function $\widehat{Z}(x,a,b,c)$, which is the shape of the solitary wave, is a heteroclinic orbit of the Hamiltonian system on $\mathbb{R}^6$,

$$\mathbf{J}_c \widehat{Z}_x = \nabla W(\widehat{Z}), \quad \widehat{Z} \in \mathbb{R}^6 \tag{2.16}$$

with

$$\mathbf{J}_c = \mathbf{K} - c\mathbf{M}, \quad W(\widehat{Z}) = S(Z_0^- + \widehat{Z}) - aP(Z_0^- + \widehat{Z}) - bQ(Z_0^- + \widehat{Z}).$$

The symplectic operator $\mathbf{J}_c$ is nondegenerate and defines the symplectic structure $(\mathbb{R}^6, \Omega)$, where $\Omega = \kappa - c\omega$.

This Hamiltonian system is the analogue of the Hamiltonian ODE presented in (1.5). There are, however, two important differences: the symplectic structure Ω is defined explicitly in terms of a combination of the spatial (κ) and temporal (ω) structures, and c appears here explicitly as a multiplier of the temporal symplectic structure ω. In other words, even though (1.5) is Hamiltonian there is no connection with the spatial or temporal symplectic structure of the full system (1.1), while (2.16) still contains these connections.

The heteroclinic orbit $\widehat{Z}(x,a,b,c)$ satisfies the asymptotic conditions

$$\lim_{x\to-\infty} \|\widehat{Z}(x,a,b,c)\| = 0 \quad \text{and} \quad \lim_{x\to\infty} \|\widehat{Z}(x,a,b,c) - Z_0^+ + Z_0^-\| = 0\,,$$

where $Z_0^+ = \mathcal{G}_\gamma Z_0^-$ for some $\gamma \in G$. In other words, as $x \to +\infty$ the function $\widehat{Z}(x,a,b,c)$ is asymptotic to a point on $\mathcal{M}_\infty$ other than Z_0^-, but this point is related to Z_0^- by an element γ in the group G. In the present case, the difference in Z_0^+ and Z_0^- corresponds to a jump in the value of the potential q_1.

3. A three-parameter family of solitary waves.

For the linearized stability theory, we will assume the existence of open sets A, B, and C in $\mathbb{R}$ such that for each $(a,b,c) \in A \times B \times C$ there exists a bounded travelling wave shape $\widetilde{Z}(x;a,b,c) = Z_0^-(a,b,c) + \widehat{Z}(x;a,b,c)$, which satisfies

$$\mathbf{J}_c \widetilde{Z}_x = \nabla S(\widetilde{Z}) - a\nabla P(\widetilde{Z}) - b\nabla Q(\widetilde{Z}). \tag{3.1}$$

Furthermore, we assume that the derivative of the shape of the solitary wave, $\widetilde{Z}_x$, is exponentially decaying with asymptotic estimate

$$\lim_{x\to\pm\infty} e^{\pm\delta x} \widetilde{Z}_x = \Psi^\pm \quad \text{and} \quad \lim_{x\to\pm\infty} \partial_x [e^{\pm\delta x} \widetilde{Z}_x] = 0 \tag{3.2}$$

for some $\Psi^\pm \in \mathbb{R}^6$ and $\delta > 0$. This assumption is in general easy to verify for solitary waves which are explicitly known. Indeed the above two hypotheses are very unrestrictive and cover a wide range of known solitary waves.

The bounded travelling wave shapes $\widetilde{Z}$ will be asymptotic to the points Z_0^+ and Z_0^- for $x \to \infty$, respectively, $x \to -\infty$. The phase shift between the point on $\mathcal{M}_\infty$ at plus and minus infinity are related by using the group action of G: explicitly we find

$$Z_0^+ = \mathcal{G}_\gamma Z_0^-, \quad \text{where} \quad \gamma = \int_{-\infty}^{\infty} (u(x) - b)\, dx,$$

and $u(x) = \widetilde{q}_2(x)$ is the second component of the solitary wave solution. By the hypotheses, this integral exists and is nonzero in general. It is a generalization of the "mass" of the solitary wave (cf. Longuet-Higgins [27]).

The momentum of the shape of the solitary wave is defined by

$$I(\widetilde{Z}) = \frac{1}{2}\int_{-\infty}^{+\infty} \omega(\widetilde{Z}, \widetilde{Z}_x)\, dx\,, \tag{3.3}$$

where the dependence on a and b has been suppressed. By taking

$$H(\widetilde{Z}) = \int_{-\infty}^{+\infty} (\tfrac{1}{2}\kappa(\widetilde{Z}, \widetilde{Z}_x) + W(\widetilde{Z}))\, dx\,,$$

it is straightforward to verify that the energy-momentum characterization of solitary waves is encoded in (2.16) and (3.1) (cf. [7], [9]), but this characterization will not be needed explicitly in what follows.

With the above hypotheses, it is shown in [9] that the derivative of I with respect to c exists and takes the form

$$\frac{d}{dc} I(\widetilde{Z}) = \int_{-\infty}^{+\infty} \omega(\widetilde{Z}_c, \widetilde{Z}_x)\, dx + \frac{1}{2}\omega(Z_0^+, \partial_c Z_0^+)\,. \tag{3.4}$$

An essential point to note in the interpretation of this expression is that the derivative with respect to c is taken with all other parameters fixed. While this may appear to be obvious, it is easy to be misled into thinking that a family of solutions depends on c when in fact it exists only for a single value of c. Examples of this are the explicit solutions found in [12], [14], [15], and [20] which for fixed values of the parameters in the equation exist for a single value of c and therefore $\frac{dI}{dc}$ cannot be explicitly computed. On the other hand, when a solitary wave state is known at an isolated value of c, it is not difficult to prove that it persists for a range of c values by a multisymplectic Melnikov argument. In other words, families in c generically exist, even when an explicit solution exists for a single value of c only. An example of the numerical continuation of such an isolated explicit solution can be found in [12].

4. The linearization about a family of solitary waves. To study the stability of a solitary wave $\mathcal{G}_{\theta(x,t)}[Z_0^- + \mathcal{T}_{\tau(t)}\widehat{Z}(x;a,b,c)]$ (see (2.15)), write $Z(x,t) = \mathcal{G}_{\theta(x)}[Z_0^- + \mathcal{T}_{\tau(t)}[\widehat{Z}(x;a,b,c) + \widehat{U}(x,t)]]$. Then the linearization of (2.4) about the family of solitary waves takes the form

$$\mathbf{M}\widehat{U}_t + \mathbf{J}_c\widehat{U}_x = \mathbf{B}(x;a,b,c)\widehat{U}, \tag{4.1}$$

where

$$\begin{aligned}\mathbf{B}(x;\cdot) &= D^2W(\widehat{Z}(x;\cdot))\\ &= D^2S(Z_0^- + \widehat{Z}(x;\cdot)) - aD^2P(Z_0^- + \widehat{Z}(x;\cdot)) - bD^2Q(Z_0^- + \widehat{Z}(x;\cdot))\end{aligned}$$

(cf. section 3 of [9]). With the spectral ansatz $\widehat{U}(x,t) = e^{\lambda t}U(x,\lambda)$, the system (4.1) reduces to

$$U_x = \mathbf{A}(x,\lambda)U, \quad U \in \mathbb{C}^6, \tag{4.2}$$

with

$$\mathbf{A}(x,\lambda) = \mathbf{J}_c^{-1}[\mathbf{B}(x;\cdot) - \lambda\mathbf{M}].$$

The dependence on the parameters (a, b, c) in the argument of $\mathbf{A}$ is suppressed, as they are considered fixed in the stability analysis. The matrix $\mathbf{A}(x, \lambda)$ has the following asymptotic limits:

$$\lim_{x\to\pm\infty} \mathbf{A}^\infty(\lambda) = \mathbf{J}_c^{-1}[\mathbf{B}^\infty - \lambda\mathbf{M}]\,,$$

where

$$\mathbf{B}^\infty = \lim_{x\to\pm\infty} \mathbf{B}(x;\cdot) = D^2S(Z_0) - aD^2P(Z_0) - bD^2Q(Z_0)\,,$$

with Z_0 either Z_0^- or Z_0^+. It is a consequence of the results in [9] that although $Z_0^- \neq Z_0^+$, the linearization $\mathbf{B}^\infty$ will be the same at $\pm\infty$. Explicitly, $\mathbf{B}^\infty$ is

$$\mathbf{B}^\infty = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & F_{qq}(b,0) + \frac{E_q(b,0)^2}{\beta} & F_{qr}(b,0) + \frac{E_q(b,0)E_r(b,0)}{\beta} & 1 & 0 & \frac{E_q(b,0)}{\beta} \\ 0 & F_{qr}(b,0) + \frac{E_q(b,0)E_r(b,0)}{\beta} & \alpha + F_{rr}(b,0) + \frac{E_r(b,0)^2}{\beta} & 0 & 1 & \frac{E_r(b,0)}{\beta} \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & \frac{E_q(b,0)}{\beta} & \frac{E_r(b,0)}{\beta} & 0 & 0 & \frac{1}{\beta} \end{pmatrix}.$$

The spectrum of $\mathbf{A}^\infty(\lambda)$ is defined by

$$\sigma\left(\mathbf{A}^\infty(\lambda)\right) = \{\, \mu \in \mathbb{C} \;:\; \Delta(\mu, \lambda) = 0 \,\}\,,$$

where

$$\Delta(\mu, \lambda) = \det[\mathbf{B}^\infty - \mu\mathbf{J}_c - \lambda\mathbf{M}]\,, \quad \lambda \in \Lambda\,.$$

The set $\Lambda \in \mathbb{C}$ is some subset of the right-half complex λ-plane, which will be identified later. A straightforward calculation shows that

$$\Delta(\mu, \lambda) = \mu^6 + \frac{C_1}{\beta}\mu^4 - \frac{C_2 + 2c}{\beta}\mu^2 + 2\frac{\lambda}{\beta}\mu\,, \tag{4.3}$$

where $C_1 = F_{rr}(b, 0) - 2E_q(b, 0) + \alpha$ and $C_2 = F_{qq}(b, 0)$.

This expression shows that $\mu = 0$ is an eigenvalue for the linearized system for any value of λ. The solution of the linearized equation related to this eigenvalue is independent of x and is given explicitly by

$$U = (1, 0, 0, \lambda, 0, 0)\,. \tag{4.4}$$

This zero eigenvalue and its eigenvector are reminiscent of a similar phenomenon that appears with KdV; see section 6 of [7]. It arises due to the introduction of a potential for $u(x, t)$.

We can also determine the eigenvectors associated with each of the other five μ-eigenvalues; they are

$$U_{\mathrm{ev}}(\mu, \lambda) = (1\,,\, \mu\,,\, \mu^2\,,\, -\lambda + \mu c\,,\, -\mu F_1\,,\, -\mu F_2)\,, \quad \mu \neq 0\,, \tag{4.5}$$

where

$$F_1 = F_{qr}(b, 0) + \mu(C_1 + E_q(b, 0)) + \beta\mu^3 \quad \text{and} \quad F_2 = E_q(b, 0) + \mu E_r(b, 0) - \beta\mu^2\,.$$

If $\mu \in i\mathbb{R}\backslash\{0\}$, then $\lambda \in i\mathbb{R}$. Therefore, if $\mathrm{Re}(\lambda) > 0$, the only solution in $\sigma(\mathbf{A}^\infty(\lambda)) \cap i\mathbb{R}$ is the trivial state (4.4).

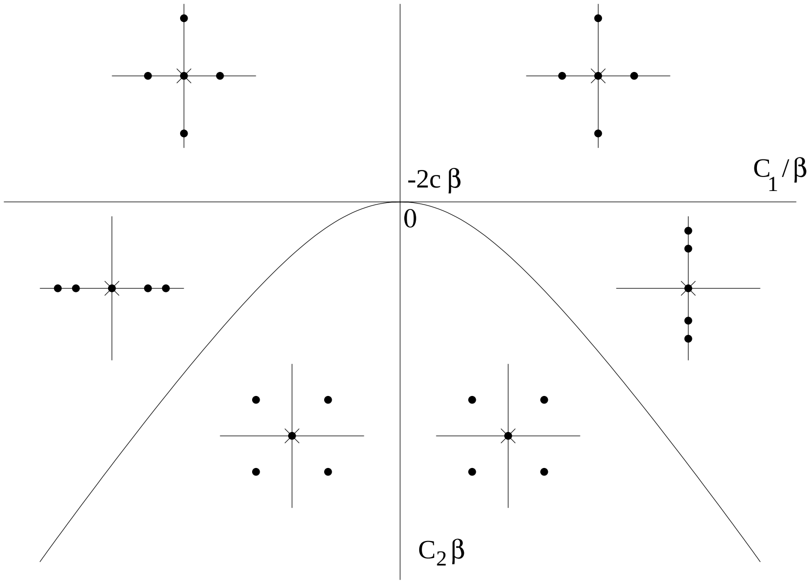


FIG. 4.1. *Sketch of the position of the eigenvalues μ at $\lambda = 0$ as a function of C_1/β and $C_2\beta$. The parabolic curve represents the relation $C_1^2 = -4\beta(C_2 + 2c)$.*

4.1. The linearized equation with $\lambda = 0$. First consider the spectrum of $\mathbf{A}^\infty(0)$ which is associated with the existing solitary wave

$$\Delta(\mu, 0) = \mu^2 \left(\mu^4 + \frac{C_1}{\beta}\mu^2 - \frac{C_2 + 2c}{\beta} \right).$$

We can immediately see that at $\lambda = 0$, the μ-spectrum is given by

$$\left\{ 0, 0, \sqrt{-\frac{C_1}{2\beta} \pm \frac{1}{2\beta}\sqrt{C_1^2 + 4\beta(C_2 + 2c)}}, -\sqrt{-\frac{C_1}{2\beta} \pm \frac{1}{2\beta}\sqrt{C_1^2 + 4\beta(C_2 + 2c)}} \right\}. \tag{4.6}$$

A sketch of the position of the eigenvalues μ as function of C_1/β and $C_2\beta$ is given in Figure 4.1. In order to satisfy the exponential decay condition (3.2) on the solitary wave, it is necessary for the spectrum $\Delta(\mu, 0)$ to have at least one pair of strictly hyperbolic eigenvalues. The region with $C_1/\beta < 0$ and $C_2\beta < -2c\beta$ with four real hyperbolic eigenvalues is the region studied by Karpman [23] using the energy-momentum method to prove stability and instability (for the case in (1.1) when f is a polynomial in u). The region with $C_1/\beta > 0$ in the neighborhood of the parabola $C_1^2 + 4\beta(C_2 + 2c) = 0$ is the region studied by Dias and Kuznetsov [16], and in this region they show that the energy-momentum method leads to the existence of a minimum (for the case in (1.1) when f is a quadratic function of u).

Here we will consider the case where the spectrum of $\mathbf{A}^\infty(0)$ has exactly one pair of hyperbolic real eigenvalues and one pair of purely imaginary eigenvalues. In Figure 4.1, this corresponds to the region

$$C_2\beta > -2c\beta\,, \tag{4.7}$$

and we will concentrate on this part of parameter space. Note that the zero eigenvalue of $\mathbf{A}^\infty(0)$ has algebraic multiplicity two and geometric multiplicity one and the

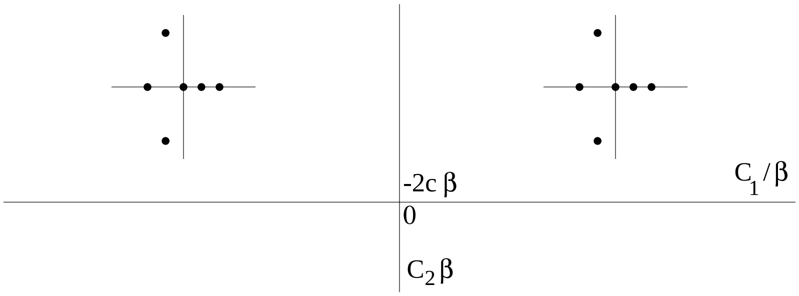


FIG. 4.2. *Position of the eigenvalues for λ small and $\beta > 0$.*

eigenvector, $(1, 0, 0, 0, 0, 0)$, is in fact the generator of the group action of G, defined in section 2.

By differentiating (3.1) with respect to x, we see that $\widetilde{Z}_x$ is a solution of the linearized equation (4.2) with $\lambda = 0$. By assumption the derivative of the solitary wave shape $\widetilde{Z}_x$ is exponentially decaying, hence both real eigenvalues can be related to the exponential decay rate of the derivative of the solitary wave shape. The negative real eigenvalue is equal to $-\delta$ and the positive real eigenvalue is equal to δ (see (3.2)), with $\Psi^\pm$ the eigenvectors of the systems at $\pm\infty$.

4.2. The linearized equation with $\Re(\lambda) > 0$ and λ small. Next we consider the linearized equation (4.2) with $\Re(\lambda) > 0$ and λ small. When λ is small, the eigenvalues which were on the imaginary axis when $\lambda = 0$ have expansions for λ small given by

$$\begin{aligned} &\mu = 0; \\ &\mu = \frac{2}{C_2 + 2c}\lambda + \mathcal{O}(\lambda^2); \\ &\mu = \pm\frac{i}{2}\sqrt{2}\sqrt{C_3} - \frac{1}{2\beta}C_3(C_3 - \frac{C_1}{\beta})\lambda + \mathcal{O}(\lambda^2), \end{aligned}$$

where $C_3 = \frac{C_1}{\beta} + \sqrt{\frac{C_1^2}{\beta^2} + \frac{4(C_2+2c)}{\beta}}$. Since $\beta(C_2 + 2c) > 0$, the term $-\frac{1}{2\beta}C_3(C_3 - \frac{C_1}{\beta})$ has sign opposite to that of $\frac{2}{C_2+2c}$. Hence, if $\beta > 0$, the nonzero eigenvalues on the imaginary axis are perturbed to the left and one of the zero eigenvalues is perturbed to the right when $\lambda \neq 0$. The position of the eigenvalues is sketched in Figure 4.2 with $\beta > 0$. If $\beta < 0$, the movement of the eigenvalues will be in the opposite direction. Hence we have a 4-2 split in the eigenvalues. This means that for $\Re(\lambda) > 0$, if $\beta < 0$, there are two eigenvalues with negative real part and four eigenvalues with nonnegative real part. And if $\beta > 0$, there are two eigenvalues with positive real part and four eigenvalues with nonpositive real part.

A straightforward calculation shows that double eigenvalues, i.e., values of λ where

$$\Delta(\mu, \lambda) = \frac{\partial}{\partial\mu}\Delta(\mu, \lambda) = 0,$$

occur at isolated real values of λ, explicitly, when

$$\lambda_0 = \pm\frac{\beta}{50}\sqrt{\frac{1}{10\beta}(C_4\beta - 3C_1)}\left(\left(C_4 - \frac{C_1}{2\beta}\right)^2 - \frac{25C_1^2}{4\beta^2}\right),$$

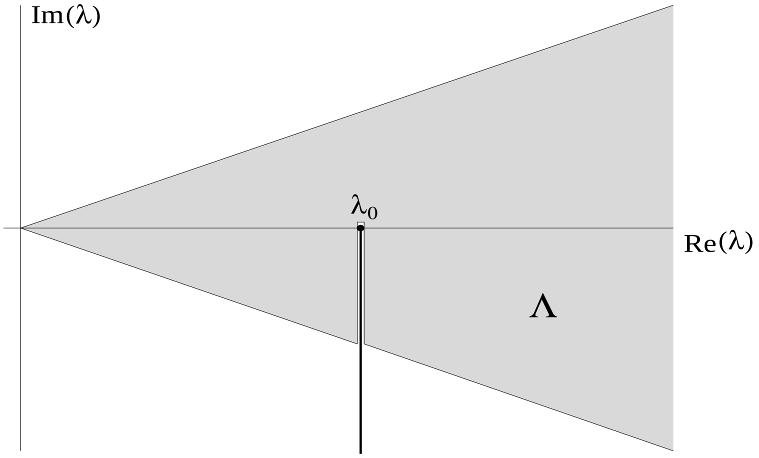


FIG. 4.3. *The set Λ_b in the right-half complex plane in which the eigenvalues are analytic and its extension is Λ.*

where $C_4 = \sqrt{\frac{9C_1^2}{\beta^2} + \frac{20(C_2+2c)}{\beta}}$. A consequence of the condition (4.7) is that λ_0 is always real. The double eigenvalues associated with $\lambda = \lambda_0$ are $\mu_d = \pm\sqrt{\frac{1}{10\beta}(C_4\beta - 3C_1)}$. If we define

$$\Lambda_b = \{\lambda \in \mathbb{C} \mid \Re(\lambda) > 0,\, |\Im(\lambda)| < \Re(\lambda)\} \backslash \{|\lambda_0| + iy \mid y \leq 0\},$$

then the eigenvalues are simple when $\lambda \in \Lambda_b$. The region Λ_b is shown in Figure 4.3.

The set Λ will be defined as the extension of the set Λ_b obtained by removing the branch point and the branch cut.

5. Intermezzo: Temporal drift along the group *G*.

In section 2, the state at infinity associated with the basic solitary wave is x and t dependent and of the form

$$Z(x,t) = \mathcal{G}_{\theta(x,t)} Z_0(a,b) = Z_0(a,b) + \theta(x,t)\, V\,. \tag{5.1}$$

Since only the first component of V is nonzero, the only component of $Z(x,t)$ in this expression which will depend on x and t is the first component. In the multisymplectic coordinates, the first coordinate is $q_1(x,t)$, where

$$u(x,t) = q_2(x,t) = \frac{\partial}{\partial x} q_1(x,t)\,, \tag{5.2}$$

and so

$$q_1(x,t) = q_1^o(a,b) + \theta(x,t) = at + bx + \theta_0\,.$$

Since $q_1(x,t)$ is a potential, the t-dependence of $q_1(x,t)$ will have no dynamic significance for $u(x,t)$ which is the primary function associated with (1.1). It is tempting to conclude that the temporal part of the flow on G—represented by the parameter a—is irrelevant. Surprisingly, it is not. By embedding the solitary wave in the three-parameter family (a,b,c), rather than the two-parameter family (b,c) (or even the one-parameter (c) family), geometric information about the linear stability is obtained, even if we set $a = 0$ at the end of the analysis. Here we will indicate two examples of how a encodes geometric information.

The basepoint, $Z_0(a, b)$, of the two-parameter family of states at infinity satisfies a constrained variational principle (cf. (2.12) and (2.13)) with the nondegeneracy condition

$$\det\begin{pmatrix} \frac{\partial P}{\partial a} & \frac{\partial P}{\partial b} \\ \frac{\partial Q}{\partial a} & \frac{\partial Q}{\partial b} \end{pmatrix} \neq 0, \tag{5.3}$$

and for the Kawahara family at infinity it was found that

$$\begin{pmatrix} \frac{\partial P}{\partial a} & \frac{\partial P}{\partial b} \\ \frac{\partial Q}{\partial a} & \frac{\partial Q}{\partial b} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -F_{qq}(b, 0) \end{pmatrix}, \tag{5.4}$$

and hence the nondegeneracy condition is satisfied for any a and b. This information was not used explicitly in the later analysis, but it does appear implicitly in the following way.

In section 4, the spectral problem associated with the linearization about the solitary wave in the limit as $x \to \pm\infty$ was associated with the function

$$\Delta(\mu, \lambda) = \det[\mathbf{B}^\infty - \mu\mathbf{J}_c - \lambda\mathbf{M}], \quad \lambda \in \Lambda. \tag{5.5}$$

In [9] the following remarkable result is proved for any relative equilibrium at infinity of the abstract form (5.1) (see Lemma 7 in [9]):

$$\Delta(\mu, \lambda) = \mathbb{C}\left[\frac{\partial Q}{\partial b}\mu^2 + \left(\frac{\partial Q}{\partial a} + \frac{\partial P}{\partial b}\right)\mu(\lambda - c\mu) + \frac{\partial P}{\partial a}(\lambda - c\mu)^2\right] + o((|\lambda| + |\mu|)^2), \tag{5.6}$$

where $\mathbb{C}$ represent a nonzero constant. In other words, *the perturbation of the μ-roots for $|\lambda|$ small is dictated by the parameter structure encoded in the state at infinity.* Substituting the Kawahara expressions into this expansion results in

$$\Delta(\mu, \lambda) = \mathbb{C}\left[-C_2\mu^2 + 2\mu(\lambda - c\mu)\right] + o((|\lambda| + |\mu|)^2), \tag{5.7}$$

using $C_2 = F_{qq}(b, 0)$. It is evident from this expression that for $|\lambda|$ small the two roots are $\mu = 0$ and

$$\mu = \frac{2}{2c + C_2}\lambda.$$

For $2c + C_2 > 0$ this result recovers precisely the perturbation result in Figure 4.2. Moreover, it gives a precise geometric description of how the zero μ-roots are perturbed when λ is perturbed away from zero for the other regions in Figure 4.1, and this information is an essential part of the construction of the Evans matrix.

A second example where the parameter a has implications is when deducing a geometric instability criterion. Using the symplectic Evans matrix, we will prove a geometric instability condition for solitary waves, based on the theory in [9], where the proof uses in an essential way the parameter structure of the state at infinity, even when only the case $a = b = 0$ is of interest.

6. The symplectic Evans matrix. The system (4.2) with the spectrum of $\mathbf{A}^\infty(\lambda)$ as shown in Figure 4.2 is in the appropriate form for construction of the Evans function. The Evans function is constructed as follows (cf. Alexander, Gardner, and Jones [1]).

For fixed $(a,b,c) \in A \times B \times C$ let $\mathbf{U}^-(x,\lambda) \in \bigwedge^2(\mathbb{C}^6)$, and let $\mathbf{U}^+(x,\lambda) \in \bigwedge^4(\mathbb{C}^6)$. Let $\alpha_-(\lambda)$ be the sum of the eigenvalues of $\mathbf{A}^\infty(\lambda)$ with positive real part, and let $\alpha_+(\lambda) = \tau_\infty(\lambda) - \alpha_-(\lambda)$, where $\tau(x,\lambda) = \mathrm{Trace}(\mathbf{A}(x,\lambda))$ and $\tau_\infty(\lambda) = \lim_{|x|\to\infty} \tau(x,\lambda)$. Then $\mathbf{U}^-(x,\lambda), \mathbf{U}^+(x,\lambda)$ are chosen to satisfy induced equations on $\bigwedge^2(\mathbb{C}^6)$ and $\bigwedge^4(\mathbb{C}^6)$, respectively, and

$$\lim_{x\to\pm\infty} \mathrm{e}^{-\alpha_\pm(\lambda)x}\mathbf{U}^\pm(x,\lambda) = \zeta_\pm(\lambda)\,,$$

where $\zeta_-(\lambda)$ and $\zeta_+(\lambda)$ are eigenvectors of $\bigwedge^2(\mathbf{A}_\infty(\lambda))$ and $\bigwedge^4(\mathbf{A}_\infty(\lambda))$, respectively, corresponding to the eigenvalues $\alpha_-(\lambda)$ and $\alpha_+(\lambda)$. The Evans function then takes the form

$$D(\lambda) = \mathrm{e}^{-\int_0^x \tau(s,\lambda)\,\mathrm{d}s}\mathbf{U}^+(x,\lambda) \wedge \mathbf{U}^-(x,\lambda) \quad \text{for all } \lambda \in \Lambda\,. \tag{6.1}$$

It is independent of x and analytic for all $\lambda \in \Lambda$ [1]. Indeed it is analytic on a larger subset of $\mathbb{C}$, but this extension will not be needed here.

One of the shortcomings of the form (6.1) is that it does not encode in any obvious way the multisymplectic structure of the system (4.2). However, by using individual solutions of (4.2), the symplectic structure can be made explicit.

For fixed $(a,b,c) \in A \times B \times C$ let $U_i^-(x,\lambda)$ for $i = 1,2$ be independent solutions of (4.2) which decay exponentially as $x \to -\infty$, and let $W_i^+(x,\lambda)$ for $i = 1,2$ be such that $\mathbf{J}_c\overline{W_i^+}$ are independent solutions of the adjoint of (4.2) which decay exponentially as $x \to +\infty$. For $\lambda \in \Lambda_b$, where Λ_b is the subset of Λ where individual vector-valued solutions are analytic, the *symplectic Evans matrix* is defined in [9] by

$$\mathbf{E}_b(\lambda) = \begin{pmatrix} \Omega(W_1^+(x,\lambda), U_1^-(x,\lambda)) & \Omega(W_1^+(x,\lambda), U_2^-(x,\lambda)) \\ \Omega(W_2^+(x,\lambda), U_1^-(x,\lambda)) & \Omega(W_2^+(x,\lambda), U_2^-(x,\lambda)) \end{pmatrix}, \tag{6.2}$$

where $\Omega(\cdot,\cdot)$ is the symplectic form associated with the Hamiltonian system (2.16). The symplectic Evans *function* is then the determinant of this matrix. If there exists a $\lambda \in \Lambda_b$ with $D_b(\lambda) = \det(\mathbf{E}_b(\lambda)) = 0$, then the basic solitary wave is linearly unstable. On the set Λ_b, $D_b(\lambda)$ and $D(\lambda)$ have the same zeros [9].

There is yet another form of the Evans function which uses individual vectors as in (6.2) but is analytic on the larger set Λ. This extension of the symplectic Evans matrix is introduced in Bridges and Derks [10]. It has the same form as (6.2) and the individual vectors in it span the same space as the vectors in (6.2) but they extend to analytic functions on the larger set Λ. Denote this Evans matrix by $\mathbf{E}(\lambda)$. In [10] it is proved that this matrix is analytic on Λ and $\det(\mathbf{E}(\lambda))$ is equal to $D(\lambda)$ on Λ. Moreover, the sign of the first nonzero derivative of $\det(\mathbf{E}_b(\lambda))$ is equal to the sign of the first nonzero derivative of $\det(\mathbf{E}(\lambda))$ at the origin.

In summary, the three forms of the Evans function can be used together to analyze the stability problem. The strategy here will be to show that $D(\lambda) = \det(\mathbf{E}(\lambda)) \to 1$ for $\lambda \to +\infty$ along the real axis. The geometry encoded in (4.2) will then be used to obtain explicit expressions for the derivatives of $D_b(\lambda) = \det(\mathbf{E}_b(\lambda))$ at the origin following [9]. Then the equivalence between $D_b(\lambda)$ and $D(\lambda)$ in Λ_b established in [10] is then used to prove a geometric instability condition.

In section 8, we will show that, for the region in parameter space associated with $C_2\beta > -2c\beta$, an explicit geometric condition for the existence of at least one unstable eigenvalue can be deduced.

7. Large $\boldsymbol{\lambda}$ behavior of the Evans function. In this section, we will prove that $D(\lambda) \to 1$ as $\lambda \to +\infty$ along the real axis, for the Evans function associated with the Kawahara equation, linearized about a solitary wave. This will be proved by applying the Pego–Weinstein lemma in the appendix to the primary form of the Evans function (6.1) on wedge spaces.

The linear system has $n = 6$ and $k = 2$, hence the wedge space $\bigwedge^2(\mathbb{C}^6)$ has dimension $d = \binom{6}{2} = 15$. Use the standard basis for $\mathbb{C}^6$ ($\mathbf{e}_1 = (1,0,0,0,0,0)^T$, etc.) and the standard lexically ordered induced basis, $\omega_1 = \mathbf{e}_1 \wedge \mathbf{e}_2, \ldots, \omega_{15} = \mathbf{e}_5 \wedge \mathbf{e}_6$.

Let $\kappa = \lambda^{-1/5}$. For λ large, the eigenvalues of the matrix $\mathbf{A}^\infty(\lambda)$ are

$$0,\ B\kappa^{-1} + \mathcal{O}(\kappa),\ \frac{1}{4}\left(-1 \pm \sqrt{5} + \mathrm{i}\sqrt{2}\sqrt{5 \pm \sqrt{5}}\,\right) B\kappa^{-1} + \mathcal{O}(\kappa)\,,$$

where $B = \sqrt[5]{\frac{2}{-\beta}}$, and

$$\frac{1}{4}\left(-1 \pm \sqrt{5} - \mathrm{i}\sqrt{2}\sqrt{5 \pm \sqrt{5}}\,\right) B\kappa^{-1} + \mathcal{O}(\kappa)\,.$$

Note that none of the eigenvalues is of order λ. This property corresponds to the fact that asymptotically $\mathbf{J}_c^{-1}\mathbf{M}$ is the main matrix in $\mathbf{A}^\infty(\lambda)$. And $\mathbf{J}_c^{-1}\mathbf{M}$ has only one eigenvalue—0—and it has algebraic multiplicity 6 and geometric multiplicity 4.

The eigenvalues of the induced matrix $\mathbf{A}_\infty^{(2)}(\lambda)$ in $\bigwedge^{(2)}(\mathbb{C}^6)$ are pairwise sums of eigenvalues of $\mathbf{A}^\infty(\lambda)$. Explicit expressions will not be given, but Figure 7.1 shows qualitatively the position of these eigenvalues relative to the eigenvalues of $\mathbf{A}^\infty(\lambda)$ in the complex μ plane.

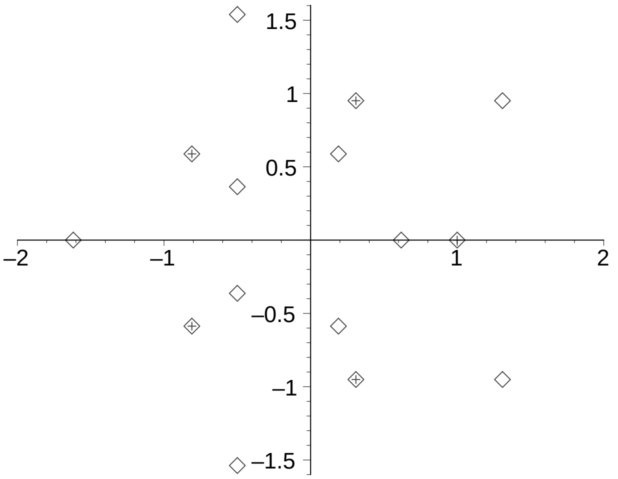


FIG. 7.1. *For $B = 1$ and λ fixed (and large), the eigenvalues of the induced matrix $\mathbf{A}_\infty^{(2)}(\lambda)$ are denoted with a diamond and the eigenvalues of the matrix $\mathbf{A}^\infty(\lambda)$ with a cross (the zero eigenvalue of $\mathbf{A}^\infty(\lambda)$ is obscured by the axes).*

It is straightforward to verify that $\mathbf{A}_\infty^{(2)}(\lambda)$ is diagonalizable for $\lambda \in \Lambda$, using the explicit expressions for the eigenvectors. The eigenvector of $\mathbf{A}^\infty(\lambda)$ for the eigenvalue 0 is $(\kappa^5, 0, 0, 1, 0, 0)^T$ and the eigenvectors for a nonzero eigenvalue μ are

$$\left(1, \mu, \mu^2, \mu c - \lambda, -\mu F_1, -\beta\mu F_2\right)^T,$$

where F_1 and F_2 are defined in (4.5). The eigenvectors of the induced matrix are wedge products of pairs of those eigenvectors, so, using the above expressions, the matrix $\mathbf{V}(\lambda)$ in the Pego–Weinstein lemma of all eigenvectors can be formed.

The columns of $\mathbf{V}(\lambda)$, the eigenvectors of the induced matrix, have the following form:

$$\begin{pmatrix} (\nu_0 - \mu_0)\kappa^8 \\ (\nu_0^2 - \mu_0^2)\kappa^7 \\ c(\nu_0 - \mu_0)\kappa^8 \\ -\beta(\nu_0^4 - \mu_0^4)\kappa^5 \\ -\beta(\mu_0^3 - \nu_0^3)\kappa^6 \\ \mu_0\nu_0(\nu_0 - \mu_0)\kappa^6 \\ (\nu_0 - \mu_0)\kappa^3 \\ -\beta\mu_0\nu_0(\nu_0^3 - \mu_0^3)\kappa^4 \\ -\beta\mu_0\nu_0(\mu_0^2 - \nu_0^2)\kappa^5 \\ (\nu_0^2 - \mu_0^2)\kappa^2 \\ -\beta\mu_0^2\nu_0^2(\nu_0^2 - \mu_0^2)\kappa^3 \\ -\beta\mu_0^2\nu_0^2(\mu_0 - \nu_0)\kappa^4 \\ -\beta(\mu_0^4 - \nu_0^4) \\ -\beta(\nu_0^3 - \mu_0^3)\kappa \\ -\beta^2\mu_0^3\nu_0^3(\nu_0 - \mu_0)\kappa^2 \end{pmatrix} + \text{h.o.t.} \quad \text{or} \quad \begin{pmatrix} -\mu_0\kappa^8 \\ -\mu_0^2\kappa^7 \\ +2\mu_0\kappa^4 \\ \mu_0^4\beta\kappa^5 \\ -\mu_0^3\beta\kappa^6 \\ 0 \\ \mu_0\kappa^3 \\ 0 \\ 0 \\ \mu_0^2\kappa^2 \\ 0 \\ 0 \\ \mu_0^4\beta \\ -\mu_0^3\beta\kappa \\ 0 \end{pmatrix} + \text{h.o.t.},$$

where μ_0 and ν_0 are two different nonvanishing solutions of the first-order approximation of the eigenvalue μ. Also, h.o.t. denotes the next order in each entry, except for the zero entries, which are identically zero.

Using this expression, we can now verify the three integral conditions in the Pego–Weinstein lemma. A straightforward but lengthy calculation gives

$$\|\mathbf{V}(\lambda)^{-1}[\mathbf{A}^{(2)}(x,\lambda) - \mathbf{A}_\infty^{(2)}(\lambda)]\mathbf{V}(\lambda)\| = \mathcal{O}(\mathrm{e}^{-\delta|x|}), \quad \text{uniform in } \lambda$$

for λ and $|x|$ large, where δ represents the exponential decay rate of the basic solitary wave. Hence, the first two integral conditions are satisfied.

For the third condition, we have in general that

$$\begin{aligned} \|\mathbf{V}(\lambda)^{-1}[\mathbf{A}^{(2)}(x,\lambda) - \mathbf{A}_\infty^{(2)}(\lambda)]\zeta_+(\lambda)\| &= \|\mathbf{V}(\lambda)^{-1}[\mathbf{A}^{(2)}(x,\lambda) - \mathbf{A}_\infty^{(2)}(\lambda)]\mathbf{V}(\lambda)\mathbf{e}_1\| \\ &= \mathcal{O}(\mathrm{e}^{-\delta|x|}). \end{aligned}$$

However, the integral of this function does not vanish for large λ, so the third integral condition of the Pego–Weinstein lemma will not be satisfied in general (for this we would require some decay rate in κ/λ in the right-hand side).

However, under the main hypothesis needed in the applications presented here—namely $E_r(b,0) = 0$—then $\mathbf{A}(x,\lambda)$ has a simpler structure and we get

$$\|\mathbf{V}(\lambda)^{-1}[\mathbf{A}^{(2)}(x,\lambda) - \mathbf{A}_\infty^{(2)}(\lambda)]\zeta_+(\lambda)\| = \mathcal{O}(\mathrm{e}^{-\delta|x|}\kappa),$$

which vanishes for large λ, and so the third integral condition is satisfied. In summary, we have proved the following. (This asymptotic result is actually true in a wedge about the real axis, but this property will not be needed here.)

Proposition 7.1. *Suppose $E_r(b,0)=0$ and $C_2\beta+2c\beta>0$, and let $D(\lambda)$ be the Evans function* (6.1) *linearized about a solitary wave of the form given in sections* 2–3*. This function satisfies $D(\lambda)\to 1$ as $\lambda\to+\infty$ along the real axis.*

8. A geometric instability criterion. Using the geometry of the symplectic Evans matrix, the slope of $\det(\mathbf{E}_b(\lambda))$ for λ small, real, and positive can be determined. Combining this result with the intermediate-value theorem and the large λ asymptotics of $D(\lambda)$, the following geometric condition for linear instability is proved in [9, 10].

Define

$$\chi_{00}^- = \left[\Omega(\Psi^-, DG_\gamma(Z_0^-)^T\Psi^+)\right]^{-1},$$

and let d_∞ be the value of the Evans function for some value of $\lambda\in\Lambda\cap\mathbb{R}$, usually λ large. Then

$$d_\infty\,\chi_{00}^-\left(\frac{\partial}{\partial c}I(\widetilde{Z})-\frac{1}{2}\omega(Z_0^+,\partial_c Z_0^+)\right)<0 \tag{8.1}$$

is a sufficient condition for linear instability of the solitary wave $\mathcal{G}_{at+bx+\theta_0}(\widetilde{Z}(x-ct))$ (see [9], [10] for full details). It follows from section 7 that $d_\infty=+1$.

The expression (8.1) can be simplified by using the properties of the existing solitary wave, the special form of $\mathbf{M}$, and the fact that DG_γ is the identity. By definition of the matrix $\mathbf{M}$, we have that

$$I(\widetilde{Z})=-\int_{-\infty}^{\infty}\widetilde{q}_2(\widetilde{q}_1)_x\,dx+\tfrac{1}{2}\widetilde{q}_2\widetilde{q}_1\big|_{-\infty}^{\infty}.$$

Using that $(\widetilde{q}_1)_x=\widetilde{q}_2-b$ and $\widetilde{q}_2(-\infty)=b=\widetilde{q}_2(\infty)$, this implies that

$$I(\widetilde{Z})=-\int_{-\infty}^{\infty}(\widetilde{q}_2-b)(\widetilde{q}_2-\tfrac{b}{2})\,dx.$$

Also,

$$\frac{1}{2}\omega(Z_0^+,\partial_c Z_0^+)=-\frac{b}{2}\frac{\partial}{\partial c}\int_{-\infty}^{\infty}(\widetilde{q}_2-b)\,dx,$$

hence

$$\frac{\partial}{\partial c}I(\widetilde{Z})-\frac{1}{2}\omega(Z_0^+,\partial_c Z_0^+)=-\frac{\partial}{\partial c}\int_{-\infty}^{\infty}(\widetilde{q}_2-b)^2\,dx.$$

Finally, since Ψ^- is an eigenvector at $\lambda=0$ with eigenvalue δ and Ψ^+ is an eigenvector at $\lambda=0$ with eigenvalue $-\delta$, there are constants $C_5^\pm$ such that $\Psi^\pm=C_5^\pm U_{\mathrm{ev}}(\mp\delta,0)$ with $U_{\mathrm{ev}}(\mu,\lambda)$ given by (4.5). Hence

$$\Psi^-=\frac{C_5^-}{C_5^+}\left(\Psi^++2\delta C_5^+(0\,,\,1\,,\,0\,,\,c\,,\,-F_{qr}(b,0)\,,\,-E_q(b,0)+\beta\delta^2)\right)$$

and

$$\left(\chi_{00}^-\right)^{-1}=-2\delta^3C_5^-C_5^+\left[\alpha+F_{rr}(b,0)-2E_q(b,0)+2\beta\delta^2\right]=-2\delta^3C_5^-C_5^+\left[C_1+2\beta\delta^2\right].$$

Since δ is the positive real eigenvalue, (4.6) gives that

$$C_1 + 2\beta\delta^2 = C_1 + \sqrt{C_1^2 + 4\beta(C_2 + 2c)} - C_1 = \sqrt{C_1^2 + 4\beta(C_2 + 2c)} > 0.$$

Also $2\delta^3 > 0$, so we can conclude the following.

THEOREM 8.1. *Define* $C_5^{\pm} = \lim_{x\to\pm\infty} e^{\pm\delta x}(\widetilde{q}_2(x) - b)$. *If* $E = 0$, $F_{qq}(b,0)\beta > -2c\beta$, *and*

$$C_5^- C_5^+ \frac{\partial}{\partial c} \int_{-\infty}^{\infty} (\widetilde{q}_2(x) - b)^2\, dx < 0,$$

then the solitary wave solution $\widetilde{Z}(x; a, b, c)$ *of the generalized Kawahara equation* (1.1) *is unstable.*

If $b = 0$ and $\widetilde{q}_2$ is even, then $C_5^- C_5^+ > 0$. In this case the condition for instability is reminiscent of the abstract condition deduced from the energy-momentum method in Grillakis, Shatah, and Strauss [18] and Bona, Souganidis, and Strauss [4], although here there is no requirement on the second variation of the constrained critical point problem. If $b \neq 0$ and $(\widetilde{q}_2 - b)$ is not even—and it is known from numerical results that such solutions exist [11]—then $C_5^- C_5^+$ may be negative, in which case the condition for instability is precisely opposite that of the energy-momentum characterization (see section 5 of [7] for an example where this switch can occur).

In the next two sections we will apply this theorem to two known classes of solitary wave states of the generalized Kawahara equation.

9. Example: Kichenasammy–Olver nonlinearity. In this section we consider a class of generalized Kawahara equations, as considered in Kichenassamy and Olver [25], i.e., $E = 0$ and $F(q, r) = -Aqr^2 + \frac{1}{2}c_1 q^2 + \frac{1}{3}c_2 q^3 + \frac{1}{4}c_3 q^4$. This implies that

$$f(u, u_x, u_{xx}) = A(u_x)^2 + 2Auu_{xx} + c_1 u + c_2 u^2 + c_3 u^3\,. \tag{9.1}$$

In [25], it is shown that if

$$c_2 = \frac{3\alpha A}{5\beta} \quad \text{and} \quad c_3 = -\frac{2A^2}{5\beta},$$

then the Kawahara equation with f given by (9.1) has a two-parameter family of exact solitary wave solutions of the form

$$\widetilde{q}_2(x) = u(x) = -\frac{10\beta\phi^2}{A}\operatorname{sech}^2(\phi(x - ct)) + b,$$

where ϕ is a positive solution of

$$80(\beta\phi^2)^2 - 20\beta\phi^2(2bA - \alpha) + 6bA(bA - \alpha) - 5\beta(c_1 + 2c) = 0, \tag{9.2}$$

which is equivalent to the condition $\Delta(\pm 2\phi, 0) = 0$. This condition implies that if $\beta(2bA - \alpha) \geq 0$, then the family exists for $40c\beta > 4b^2A^2 - 4bA\alpha - 5\alpha^2 - 20c_1\beta$. If $\beta(2bA - \alpha) < 0$, then the family exists for $40c\beta > 24b^2A^2 - 24bA\alpha - 20c_1\beta$.

In terms of the notation of section 3, we have $\delta = 2\phi$ and

$$\Psi^{\pm} = \frac{40\beta\phi^2}{A}\,(-1, \pm 2\phi, -4\phi^2, \pm 2c\phi, -4\phi^2(-4\beta\phi^2 - \alpha + 2bA), \pm 8\beta\phi^3),$$

i.e., $C_5^{\pm} = -\frac{40\beta\phi^2}{A}$. A tedious but straightforward calculation demonstrates that $\lim_{x\to\pm\infty} \partial_x[e^{\pm\delta x}\widetilde{Z}_x] = 0$. So all the conditions of section 3 are satisfied.

Next we look at section 4. For this example we have

$$C_1 = \alpha - 2Ab \quad \text{and} \quad C_2 = c_1 + \frac{6Ab}{5\beta}(\alpha - Ab).$$

So an eigenvector can be written as

$$U_{\mathrm{ev}}(\mu,\lambda) = (1\,,\,\mu\,,\,\mu^2\,,\,-\lambda+\mu c\,,\,-\mu^2(\alpha-2Ab+\beta\mu^2)\,,\,\beta\mu^3).$$

In order to have $C_2\beta > -2c\beta$, the analysis will be restricted to the case

$$40c\beta > -20\beta c_1 - 24Ab(\alpha - Ab),$$

which is exactly the sufficient condition for the existence of the family of solitary waves. This condition is also necessary if $\beta(2bA - \alpha) < 0$.

Now we are ready to apply the instability criterion. A straightforward calculation shows that

$$-\int_{-\infty}^{\infty}(u(x)-b)^2\,dx = -\frac{400\phi^3\beta^2}{3A^2}, \quad \text{hence} \quad -\frac{\partial}{\partial c}\int_{-\infty}^{\infty}(u(x)-b)^2\,dx = -\frac{400\phi^2\beta^2}{A^2}\frac{\partial}{\partial c}\phi.$$

To determine $\frac{\partial}{\partial c}\phi$, we differentiate (9.2). This gives

$$\frac{\partial}{\partial c}\phi = \frac{1}{4\phi(8\beta\phi^2 - 2bA + \alpha)}.$$

Also,

$$\left(\chi_{00}^-\right)^{-1} = (\mathbf{J}_c\Psi^-, \Psi^+) = -\frac{25600\beta^2\phi^7}{A^2}(8\beta\phi^2 + \alpha - 2bA).$$

So

$$\chi_{00}^-\left(\frac{\partial}{\partial c}I(\widetilde{Z}) - \frac{1}{2}\omega(Z_0^+, \partial_c Z_0^+)\right) = \frac{A^2}{25600\beta^2\phi^7}\frac{1}{4\phi}\frac{400\phi^2\beta^2}{A^2} = \frac{1}{256\phi^6} > 0.$$

Since the sufficient geometric condition for linear instability is not met, this suggests that the solitary wave is "stable." However, there could still be unstable λ-eigenvalues on the positive real λ-axis, but there would be two or more. The solitary wave could also be unstable due to unstable eigenvalues with nonzero imaginary part.

10. Example: Levandosky's homogeneous nonlinearity. In this section we consider a class of nonlinearities for (1.1) which includes the equations considered by Levandosky [26]. We take $E = 0$ and F homogeneous of degree $p+1$ for some $p > 1$, i.e.,

$$F(\nu q, \nu r) = \nu^{p+1}F(q,r) \quad \text{for all } \nu \geq 0 \text{ and } (q,r) \in \mathbb{R}^2.$$

In [26], it is proved that the following condition is sufficient for the existence of a one-parameter family of solitary waves,

$$\int_{-\infty}^{\infty} F(u,u_x)\,dx > 0 \quad \text{for some } u \in H^2(\mathbb{R}),\ b = 0,\ \beta = 1, \text{ and } 4c > (\max\{\alpha, 0\})^2\,.$$

Recall $\mathcal{I}(u) = \frac{1}{2}\int_{-\infty}^{\infty}|u(x)|^2\,dx$ and assume that a family of solitary waves $\widehat{u}(x;c)$ exists, where $\widehat{u}$ travels with speed c. In [26] it is also shown that under the above conditions, the solitary waves are stable if $\frac{d}{dc}\mathcal{I}(\widetilde{u}) > 0$ and unstable if $\frac{d}{dc}\mathcal{I}(\widetilde{u}) < 0$ and $p \geq 2$. When $b = 0$ in the formulation in sections 2–6, this instability condition agrees with Theorem 8.1 if $C_5^+ C_5^- > 0$.

First, we embed the solitary wave state in a two-parameter family, allowing for b to be nonzero. Then we show how our theory recovers the result in [26], without any specific use of the variational principle associated with the energy-momentum characterization. For definiteness, consider the case $\alpha = 0$ and take F to be homogeneous in r itself, i.e., there is some $0 \leq n \leq p+1$ such that

$$F(q, \nu r) = \nu^n F(q,r) \quad \text{for all } \nu \geq 0 \text{ and } (q,r) \in \mathbb{R}^2.$$

This homogeneity condition in r implies that $F_{qq}(b,0) = 0$, hence the condition on c of Theorem 8.1 becomes $c > 0$.

When there exists a family of solitary wave solutions $\{\widetilde{Z}(x;a,b,1) \mid a \in \mathbb{R},\, b \in \mathbb{R}\}$ with speed $c = 1$, then we can construct a family of solitary wave solutions with speed $c > 0$:

$$\widetilde{Z}(x;a,b,c) = c^{\frac{4-n}{4(p-1)}}\,\mathrm{diag}(c^{-\frac{1}{4}}, 1, c^{\frac{1}{4}}, c, c^{\frac{3}{4}}, c^{\frac{1}{2}})\,\widetilde{Z}(c^{\frac{1}{4}}x; ac^{\frac{n-4p}{4(p-1)}}, bc^{-\frac{4-n}{4(p-1)}}, 1).$$

Indeed, $\widetilde{Z}(x;a,b,c)$ as defined above satisfies

$$\begin{aligned}
\mathbf{J}_c D_x \widetilde{Z}(x;a,b,c) &= c^{\frac{1}{4}}\, c^{\frac{4-n}{4(p-1)}}\,\mathbf{J}_c\,\mathrm{diag}(c^{-\frac{1}{4}}, 1, c^{\frac{1}{4}}, c, c^{\frac{3}{4}}, c^{\frac{1}{2}})\,\mathbf{J}_1^{-1}\mathbf{J}_1 \\
&\qquad \widetilde{Z}_x(c^{\frac{1}{4}}x; ac^{\frac{n-4p}{4(p-1)}}, bc^{-\frac{4-n}{4(p-1)}}, 1) \\
&= c^{\frac{1}{4}}\,\mathbf{J}_c\,\mathrm{diag}(c^{-\frac{1}{4}}, 1, c^{\frac{1}{4}}, c, c^{\frac{3}{4}}, c^{\frac{1}{2}})\,\mathbf{J}_1^{-1} \\
&\qquad (\nabla S(\widetilde{Z}) - ac^{\frac{n-4p}{4(p-1)}}\nabla P(\widetilde{Z}) - bc^{-\frac{4-n}{4(p-1)}}\nabla Q(\widetilde{Z})) \\
&= \nabla S(\widetilde{Z}(x;a,b,c)) - a\nabla P(\widetilde{Z}(x;a,b,c)) - b\nabla Q(\widetilde{Z}(x;a,b,c)).
\end{aligned}$$

This implies that for $b = 0$ or $n = 4$

$$\begin{aligned}
\int_{-\infty}^{\infty}(\widetilde{q}_2(x;a,b,c) - b)^2\,dx &= c^{\frac{4-n}{2(p-1)}}\int_{-\infty}^{\infty}(\widetilde{q}_2(c^{\frac{1}{4}}x; ac^{\frac{n-4p}{4(p-1)}}, bc^{-\frac{4-n}{4(p-1)}}, 1) \\
&\qquad\qquad - bc^{-\frac{4-n}{4(p-1)}})^2\,dx \\
&= c^{\frac{4-n}{2(p-1)}}c^{-\frac{1}{4}}\int_{-\infty}^{\infty}(\widetilde{q}_2(s; ac^{\frac{n-4p}{4(p-1)}}, bc^{-\frac{4-n}{4(p-1)}}, 1) \\
&\qquad\qquad - bc^{-\frac{4-n}{4(p-1)}})^2\,ds \\
&= c^{\frac{4-n}{2(p-1)}-\frac{1}{4}}\int_{-\infty}^{\infty}(\widetilde{q}_2(s; ac^{\frac{n-4p}{4(p-1)}}, b, 1) - b)^2\,ds.
\end{aligned}$$

In the last step we used that $b = 0$ or $n = 4$, hence $bc^{-\frac{4-n}{4(p-1)}} = b$. Since the defining equation for $\widetilde{q}_2$ does not depend on a, this integral will not depend on a either and we can put $a = 0$; hence,

$$\int_{-\infty}^{\infty}(\widetilde{q}_2(x;a,b,c) - b)^2\,dx = c^{\frac{9-2n-p}{4(p-1)}}\int_{-\infty}^{\infty}(\widetilde{q}_2(s;0,b,1) - b)^2\,ds.$$

So

$$\frac{\partial}{\partial c}\int_{-\infty}^{\infty}(\widetilde{q}_2(x;a,b,c)-b)^2\,dx=\frac{9-2n-p}{4(p-1)}\,c^{\frac{13-2n-5p}{4(p-1)}}\int_{-\infty}^{\infty}(\widetilde{q}_2(s;0,b,1)-b)^2\,ds.$$

Furthermore,

$$C_5^+(a,b,c)\,C_5^-(a,b,c)=c^{\frac{4-n}{2(p-1)}}C_5^+(ac^{\frac{n-4p}{4(p-1)}},b,1)\,C_5^-(ac^{\frac{n-4p}{4(p-1)}},b,1).$$

It is difficult to verify the hypothesis (3.2) for this example and therefore we assume (3.2) is satisfied. Then with

$$C_5^+(ac^{\frac{n-4p}{4(p-1)}},b,1)\,C_5^-(ac^{\frac{n-4p}{4(p-1)}},b,1)\int_{-\infty}^{\infty}\widetilde{q}_2^2(s;0,b,1)\,ds>0.$$

Theorem 8.1 shows that the solitary waves are unstable if $9-2n-p<0$, i.e., $p>9-2n$. These instability results agree with those in [26] when an error in Lemma 3.3 in [26] is corrected. (The $(p+1)(4-\beta)$ in the numerator of the expression for γ should be replaced by $4(p+1)-2\beta$.)

If $n=4$, then $b\neq 0$ is allowed and we obtain that the wave is always unstable since

$$\frac{\partial}{\partial c}\int_{-\infty}^{\infty}(\widetilde{q}_2(x;a,b,c)-b)^2\,dx=-\frac{1}{4}c^{-\frac{5}{4}}\int_{-\infty}^{\infty}\widetilde{q}_2^2(s;0,b,1)\,ds<0.$$

Appendix. Large $\boldsymbol{\lambda}$ behavior and the Pego–Weinstein lemma. The following result is a generalization of Proposition 1.17 in Pego and Weinstein [29], which gives a sufficient condition for $D(\lambda)\to 1$ as $\lambda\to+\infty$ along the real λ-axis. Consider the system

$$\mathbf{u}_x=\mathbf{A}(x,\lambda)\mathbf{u}\,,\quad \mathbf{u}\in\mathbb{C}^n\,,\quad \lambda\in\Lambda\,, \tag{A.1}$$

where Λ is an open simply connected subset of $\mathbb{C}$, and Λ includes a wedge about the real axis in which we can take $|\lambda|\to\infty$. The spectrum of $\mathbf{A}^\infty(\lambda)$, where

$$\mathbf{A}^\infty(\lambda)=\lim_{|x|\to\infty}\mathbf{A}(x,\lambda), \tag{A.2}$$

is assumed to have k-eigenvalues with negative real part and $n-k$ with nonnegative real part. A critical hypothesis in Proposition 1.17 in [29] is that $k=1$. However, this hypothesis is not essential if we take into account that on $\bigwedge^k(\mathbb{C}^n)$ the induced matrix

$$\mathbf{A}_\infty^{(k)}(\lambda)\stackrel{\text{def}}{=}\bigwedge\nolimits^k(\mathbf{A}^\infty(\lambda))$$

has a unique simple eigenvalue of largest negative real part. Then working on $\bigwedge^k(\mathbb{C}^n)$ with the Evans function also on the exterior algebra, the proof of Proposition 1.17 carries over [10]. The precise statement of the result needed in this paper is given below. Although stated in a substantially more general form, the proof given in [29] carries over almost verbatim.

Pego–Weinstein lemma. *Consider the system* (A.1)–(A.2) *and suppose that for all $\lambda\in\Lambda$ the eigenvalue of $\mathbf{A}_\infty^{(k)}(\lambda)$ with largest negative real part is unique and simple. Denote this eigenvalue by $\alpha(\lambda)$ and its (analytic choice of) right eigenvector by $\zeta(\lambda)$*

and its (analytic choice of) left eigenvector by $\eta(\lambda)$ *with normalization* $[\![\eta,\zeta]\!]_k = 1$, *where* $[\![\cdot,\cdot]\!]_k$ *is the induced inner product on* $\bigwedge^k(\mathbb{C}^n)$.

Let $\mathbf{U}(x,\lambda) \in \bigwedge^k(\mathbb{C}^n)$ *be the solution of the system*

$$\mathbf{U}_x = \mathbf{A}^{(k)}(x,\lambda)\mathbf{U} \quad \textit{satisfying} \quad \lim_{x\to+\infty} \mathrm{e}^{-\alpha(\lambda)x}\mathbf{U}(x,\lambda) = \zeta(\lambda) \in \textstyle\bigwedge^k(\mathbb{C}^n)\,.$$

Similarly, let $\mathbf{W}(x,\lambda) \in \bigwedge^k(\mathbb{C}^n)$ *be the solution of the system*

$$\mathbf{W}_x = -\mathbf{A}^{(k)}(x,\lambda)^T\mathbf{W} \quad \textit{satisfying} \quad \lim_{x\to-\infty} \mathrm{e}^{\alpha(\lambda)x}\mathbf{W}(x,\lambda) = \eta(\lambda) \in \textstyle\bigwedge^k(\mathbb{C}^n)\,.$$

In terms of these functions, the Evans function (6.1) *can be expressed in the form*

$$D(\lambda) = \mathbf{W}(0,\lambda)\cdot\mathbf{U}(0,\lambda) \overset{\text{def}}{=} \langle\overline{\mathbf{W}(0,\lambda)},\mathbf{U}(0,\lambda)\rangle,$$

where $\langle\cdot,\cdot\rangle$ *is a standard Hermitian inner product on* $\mathbb{C}^d$ *and* $d = \dim\bigwedge^k(\mathbb{C}^n)$.

Now, suppose $\mathbf{A}_\infty^{(k)}(\lambda)$ *is diagonalizable for large* λ *and let* $\mathbf{V}(\lambda) \in \mathbb{C}^{d\times d}$ *be the matrix of right eigenvectors such that the first column is* $\zeta(\lambda)$. *If*

$$\int_{-\infty}^{+\infty} \|\mathbf{V}(\lambda)^{-1}[\mathbf{A}^{(k)}(x,\lambda) - \mathbf{A}_\infty^{(k)}(\lambda)]\mathbf{V}(\lambda)\|\,\mathrm{d}x \le C\,, \quad \textit{independent of } \lambda,$$

$$\int_{|x|\ge x_0} \|\mathbf{V}(\lambda)^{-1}[\mathbf{A}^{(k)}(x,\lambda) - \mathbf{A}_\infty^{(k)}(\lambda)]\mathbf{V}(\lambda)\|\,\mathrm{d}x \to 0\,, \quad \textit{as } x_0\to\infty\textit{, uniformly in } \lambda,$$

$$\int_{-\infty}^{+\infty} \|\mathbf{V}(\lambda)^{-1}[\mathbf{A}^{(k)}(x,\lambda) - \mathbf{A}_\infty^{(k)}(\lambda)]\zeta(\lambda)\|\,\mathrm{d}x \to 0\,, \quad \textit{as } |\lambda|\to\infty\,,$$

then

$$\mathbf{V}(\lambda)^{-1}\mathbf{U}(0,\lambda) = \mathbf{V}(\lambda)^{-1}\zeta(\lambda) + o(1) \quad \textit{for } |\lambda|\to\infty \tag{A.3}$$

and $\mathbf{W}(0,\lambda)\mathbf{V}(\lambda)$ *is bounded with*

$$\mathbf{W}(0,\lambda)\mathbf{V}(\lambda)\mathbf{e}_1 = \mathbf{W}(0,\lambda)\zeta(\lambda) = 1 + o(1) \quad \textit{for } |\lambda|\to\infty\,. \tag{A.4}$$

The two results (A.3) *and* (A.4) *imply that* $D(\lambda)\to 1$ *as* $|\lambda|\to\infty$.

REFERENCES

[1] J. Alexander, R. Gardner, and C.K.R.T. Jones, *A topological invariant arising in the stability analysis of traveling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.

[2] T.B. Benjamin, *The stability of solitary waves*, Proc. Roy. Soc. (London) Ser. A, 328 (1972), pp. 153–183.

[3] J.L. Bona, *On the stability of solitary waves*, Proc. Roy. Soc. (London) Ser. A, 344 (1975), pp. 363–374.

[4] J.L. Bona, P.E. Souganidis, and W.A. Strauss, *Stability and instability of solitary waves of KdV type*, Proc. Roy. Soc. (London) Ser. A, 411 (1987), pp. 395–411.

[5] T.J. Bridges, *Multi-symplectic structures and wave propagation*, Math. Proc. Cambridge Philos. Soc., 121 (1997), pp. 147–190.

[6] T.J. Bridges, *Toral equivariant partial differential equations and quasiperiodic patterns*, Nonlinearity, 11 (1998), pp. 467–500.

[7] T.J. Bridges and G. Derks, *Unstable eigenvalues, and the linearization about solitary waves and fronts with symmetry*, Proc. Roy. Soc. (London) Ser. A, 455 (1999), pp. 2427–2469.

[8] T.J. Bridges and G. Derks, *Hodge duality and the Evans function*, Phys. Lett. A, 251 (1999), pp. 363–372.

[9] T.J. Bridges and G. Derks, *The symplectic Evans matrix, and the instability of solitary waves and fronts with symmetry*, Arch. Ration. Mech. Anal., 156 (2001), pp. 1–87.

[10] T.J. Bridges and G. Derks, *Constructing the Symplectic Evans Matrix Using Maximally Analytic Individual Vectors*, preprint, University of Surrey, Surrey, UK, 2001; also available online from http://www.maths.surrey.ac.uk/personal/st/T.Bridges/PAPERS/SEMpaper.ps.

[11] A.R. Champneys, *Homoclinic orbits in reversible systems and their applications in mechanics, fluids and optics*, Phys. D, 112 (1998), pp. 158–186.

[12] A.R. Champneys and M.D. Groves, *A global investigation of solitary-wave solutions to a two-parameter model for water waves*, J. Fluid Mech., 342 (1997), pp. 199–229.

[13] W. Craig and M.D. Groves *Hamiltonian long-wave approximations to the water-wave problem*, Wave Motion, 19 (1994), pp. 367–389.

[14] X. Dai and J. Dai, *Some solitary wave solutions for families of generalized higher-order KdV equations*, Phys. Lett. A, 142 (1989), pp. 367–370.

[15] B. Dey, A. Khare, and C.N. Kumar, *Stationary solutions of the fifth-order KdV-type equations and their stabilization*, Phys. Lett. A, 223 (1996), pp. 449–452.

[16] F. Dias and E.A. Kuznetsov, *Nonlinear stability of solitons in the fifth-order Korteweg-de Vries equation*, Phys. Lett. A, 263 (1999), pp. 98–104.

[17] J. W. Evans, *Nerve axon equations* IV. *The stable and unstable impulse*, Indiana Univ. Math. J., 24 (1975), pp. 1169–1190.

[18] M. Grillakis, J. Shatah, and W. Strauss, *Stability theory of solitary waves in the presence of symmetry,* I, J. Funct. Anal., 74 (1987), pp. 160–197.

[19] M. Groves, *Solitary-wave solutions to a class of fifth-order model equations*, Nonlinearity, 11 (1998), pp. 341–353.

[20] G. Huang, S. Luo, and X. Dai, *Exact and explicit solitary-wave solutions to a model equation for water waves*, Phys. Lett. A, 139 (1989), pp. 373–374.

[21] A.T. Ill'ichev and A.Y. Semenov, *Stability of solitary waves in dispersive media described by a fifth-order evolution equation*, Theor. Comput. Fluid Dyn., 3 (1992), pp. 307–326.

[22] B.B. Kadomtsev and V.I. Petviashvili, *On the stability of solitary waves in weakly dispersing media*, Sov. Phys. Dokl., 15 (1970), pp. 539–541.

[23] V.I. Karpman, *Stabilization of soliton instabilities by higher-order dispersion: KdV-type equations*, Phys. Lett. A, 210 (1996), pp. 77–84.

[24] R. Kawahara, *Oscillatory solitary waves in dispersive media*, J. Phys. Soc. Japan, 33 (1972), pp. 260–264.

[25] S. Kichenassamy and P.J. Olver, *Existence and nonexistence of solitary wave solutions to higher-order model evolution equations*, SIAM J. Math. Anal., 23 (1992), pp. 1141–1166.

[26] S.P. Levandosky, *A stability analysis for fifth-order water-wave models*, Phys. D, 125 (1999), pp. 222–240.

[27] M.S. Longuet-Higgins, *On the mass, momentum, energy and circulation of a solitary wave*, Proc. Roy. Soc. (London) Ser. A, 337 (1974), pp. 1–37.

[28] P.J. Olver, *Hamiltonian perturbation theory and water waves*, Contemp. Math., 28 (1984), pp. 231–249.

[29] 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.

SIAM J. MATH. ANAL.
Vol. 33, No. 6, pp. 1379–1392

NONEXISTENCE OF PERMANENT CURRENTS IN CONVEX PLANAR SAMPLES*

SHUICHI JIMBO† AND PETER STERNBERG‡

Abstract. Recent works have demonstrated the existence of nontrivial stable critical points of the Ginzburg–Landau energy

$$(\Psi, A) \to \int_\Omega \frac{1}{2}|(\nabla - iA)\Psi|^2 + \frac{\kappa^2}{4}(1-|\Psi|^2)^2\, dx + \frac{1}{2}\int_{\mathbb{R}^n} |\mathrm{curl}|\, \mathrm{A}^2\, dx$$

for multiply connected domains $\Omega \subset \mathbb{R}^n$ with $n = 2$ or 3 and for simply connected domains Ω that are close in L^1 to multiply connected domains. In this article we demonstrate that while there is no topological obstruction to the presence of such stable critical points, there is a geometric obstruction. Specifically, we show the nonexistence of stable critical points of this energy in two-dimensional convex domains.

Key words. Ginzburg–Landau system, stability, convex domains

AMS subject classifications. 35J50, 35J60, 49K40

PII. S0036141001386027

1. Introduction.

The phenomenon of permanent currents in superconducting materials has been well known in the physics community for many years. In a typical experiment, a ring-shaped sample is subjected to an applied field that induces a current. The temperature of the sample is then lowered into the superconducting regime, the applied field is shut off, and the current persists with only negligible resistance—sometimes for a period of years. Mathematically, one can describe these states using the Ginzburg–Landau theory of superconductivity [3, 5]. Within this theory, a permanent current in two dimensions corresponds to a nontrivial stable critical point of the energy

$$G(\Psi, A) = \int_\Omega \frac{1}{2}\left|(\nabla - iA)\Psi\right|^2 + \frac{\kappa^2}{4}(1-|\Psi|^2)^2\, dx + \frac{1}{2}\int_{\mathbb{R}^2} \left|\mathrm{curl}\, A\right|^2\, dx.$$

Here $\Omega \subset \mathbb{R}^2$ is a bounded cross-section of a cylindrical sample such as a wire or thin film. The magnitude of the order parameter $\Psi : \Omega \to \mathbb{C}$ measures the density of superconducting electron pairs, $A : \mathbb{R}^2 \to \mathbb{R}^2$ is the magnetic potential whose curl represents the induced magnetic field, and κ is the Ginzburg–Landau parameter. The energy G is most naturally defined for pairs (Ψ, A) having square-integrable derivatives so we consider it as defined on the space $H^1(\Omega; \mathbb{C}) \times Z$ where

$$Z \equiv \{A \in H^1_{\mathrm{loc}}(\mathbb{R}^2; \mathbb{R}^2) : \mathrm{curl}\, A \in L^2(\mathbb{R}^2; \mathbb{R}^2)\}. \tag{1.1}$$

We recall that the Ginzburg–Landau energy enjoys the gauge invariance property that

$$G(\Psi, A) = G(\Psi e^{i\phi}, A + \nabla\phi) \tag{1.2}$$

*Received by the editors March 2, 2001; accepted for publication (in revised form) November 19, 2001; published electronically May 23, 2002.
http://www.siam.org/journals/sima/33-6/38602.html

†Department of Mathematics, Hokkaido University, Sapporo 060, Japan (jimbo@math.sci.hokudai.ac.jp). The research of this author was supported by Grant-in-Aid for Scientific Research 09440045, Ministry of Education, Culture, and Science, Japan.

‡Department of Mathematics, Indiana University, Bloomington, IN 47405 (sternber@indiana.edu). The research of this author was partially supported by NSF grant DMS-9322617.

for any $\phi \in H^2_{loc}(\mathbb{R}^2)$. In section 2 we will discuss the gauge choices made in this paper.

Any pair (Ψ, A) in $H^1(\Omega;\mathbb{C}) \times Z$ for which the first variation of G vanishes will be referred to as a critical point. By a stable critical point, we will mean a critical point for which the second variation is nonnegative. (See the beginning of section 2 for the precise definitions.) By a nontrivial critical point we will mean one that is not gauge-equivalent to a state $(c, 0)$ for $c \in \mathbb{C}$.

Recently, using a degree theoretic approach, we have succeeded in proving the existence of these stable critical points when κ^2 is sufficiently large in both two and three dimensions, provided the sample is multiply connected [6, 7, 10]. Indeed, one finds, roughly speaking, that there exists a stable critical point living in each homotopy class of mappings from Ω into S^1. One may assume that these solutions correspond to the experimentally observed permanent currents produced in ring-shaped samples in the laboratory.

Perhaps more surprising, however, is the more recent discovery in [8] that stable critical points exist in certain simply connected domains, thus showing that there is no topological obstruction to the presence of permanent currents. These critical points are constructed through a perturbation of domain argument, starting from a multiply connected domain. In particular, then, they are shown to exist in highly nonconvex samples. They differ from the ones found in multiply connected domains in that they contain vortices—that is, zeros of the order parameter—unlike the vortex-free solutions of [6, 7, 10].

In light of these existence results, we pursue here the question of whether there are any geometric obstructions to producing permanent currents via Ginzburg–Landau theory. That is, we ask whether among simply connected domains Ω, there are any conditions on $\partial\Omega$ that would preclude the existence of nontrivial stable critical points. We should perhaps point out that if we drop the requirement of stability in our question, then there certainly do exist nontrivial critical points even under very stringent assumptions on the geometry of the sample. For example, taking $\Omega \subset \mathbb{R}^2$ to be a disc, one can construct critical points (Ψ, A) under the radial ansatz

$$\Psi = \Psi(r,\theta) = w_m(r)e^{im\theta}, \quad A = A(r,\theta) = \frac{Y_m(r)}{r}(-\sin\theta, \cos\theta)$$

for each positive integer m, where then w_m and Y_m are obtained through minimization of G over this class of radial competitors.

It turns out, however, that all of these radial solutions in a disc are unstable. This follows as a special case of our main result, stated below, which rules out the existence of nontrivial stable critical points of G in two dimensions, whenever Ω is convex.

THEOREM 1.1. *Let $\Omega \subset \mathbb{R}^2$ be an open, bounded convex set with $C^{5,\alpha}$ boundary, for any $\alpha \in (0,1)$. Then the only stable critical point of G is the pair* $(1,0)$ *or one of its gauge-equivalent representations.*

Our approach should be viewed as a descendant of the technique employed in [2, 9] to prove the analogous result for the well-studied scalar functional

$$u \to \int_\Omega F(u) + |\nabla u|^2 \, dx$$

for $F : \mathbb{R}^1 \to \mathbb{R}^1$, $u : \Omega \to \mathbb{R}^1$, and $\Omega \subset \mathbb{R}^n$ convex. It is based upon an explicit expression for the second variation of G computed about a critical point. In Proposition 3.1, we show that we can describe this second variation completely in terms of boundary integrals that include a dependence on the curvature of $\partial\Omega$. Under the convexity assumption, we then show through Proposition 3.2 and Theorem 3.3 that the second variation is negative unless $(\Psi, A) = (1, 0)$ (or any of its gauge-equivalent representations).

The paper is organized as follows. In section 2 we review basic properties of the Ginzburg–Landau system and develop the regularity theory needed for our approach. We should note that we have not attempted to prove an optimal theorem with regard to smoothness of the boundary. The theorem above contains an assumption that $\partial\Omega \in C^{5,\alpha}$, meaning that, locally, the boundary can be described as the graph of a function having five continuous derivatives with fifth derivatives satisfying a Hölder condition with exponent α; cf. [4]. This assumption can surely be relaxed. In section 2 we address the reason for the assumed regularity. We also present a simple lemma relating curl-free vector fields satisfying a Neumann boundary condition to the curvature of the boundary. Then in section 3 we work with the second variation of the Ginzburg–Landau energy in order to obtain our main result.

2. Preliminaries. Throughout this paper we shall take $\Omega \subset \mathbb{R}^2$ to be a bounded, open, convex set with sufficiently smooth boundary, say $\partial\Omega \in C^{5,\alpha}$. We will denote the outer unit normal to $\partial\Omega$ by ν and the mean curvature of $\partial\Omega$ by H so that $H(x) \geq 0$ for all $x \in \partial\Omega$. We denote a ball centered at x of radius R by $B(x, R)$ and we use Ψ^* to denote the complex conjugate of a complex-valued function Ψ. We will frequently invoke the summation convention on expressions with repeated indices.

Before discussing the nature of critical points, we discuss the two gauge choices we will utilize in our approach. The first is to work only with vector potentials $A \in Z$ (cf. (1.1)) satisfying the condition

$$\operatorname{div} A = 0 \text{ in } \mathbb{R}^2. \tag{2.1}$$

This choice is clearly possible in light of the solvability in $H^2_{\text{loc}}(\mathbb{R}^2)$ of Poisson's equation

$$-\Delta\phi = \operatorname{div} A \in L^2_{\text{loc}}(\mathbb{R}^2).$$

A second choice we will use is the following:

$$\operatorname{div} A = 0 \text{ in } \Omega, \quad A \cdot \nu = 0 \text{ on } \partial\Omega. \tag{2.2}$$

Such a gauge is obtainable via the transformation $\Psi \to \Psi e^{i\phi}$, $A \to A + \nabla\phi$ where $\phi \in H^2_{\text{loc}}(\mathbb{R}^2)$ is taken to be an extension to $\mathbb{R}^2$ of the solution to the boundary value problem

$$\Delta\phi = -\operatorname{div} A \text{ in } \Omega, \quad \nabla\phi \cdot \nu = -A \cdot \nu \text{ on } \partial\Omega.$$

Of course, when working in this gauge we have not uniquely determined A since the extension of ϕ to the complement of Ω is arbitrary, but we make here the convention that we shall always take as smooth an extension as possible, depending on the known regularity of A.

By a critical point of G we shall mean a pair $(\Psi, A) \in H^1(\Omega; \mathbb{C}) \times Z$ such that

$$\frac{d}{d\varepsilon} G(\Psi + \varepsilon\tilde{\Psi}, A + \varepsilon\tilde{A})_{|_{\varepsilon=0}} = 0 \quad \text{for all } (\tilde{\Psi}, \tilde{A}) \in H^1(\Omega; \mathbb{C}) \times Z. \tag{2.3}$$

One readily verifies that such a critical point will satisfy the Ginzburg–Landau system

$$(\nabla - iA)^2\Psi + \kappa^2(1-|\Psi|^2)\Psi = 0 \quad \text{in } \Omega, \tag{2.4}$$

$$\operatorname{curl}\operatorname{curl} A + \left\{\frac{i}{2}(\Psi^*\nabla\Psi - \Psi\nabla\Psi^*) + |\Psi|^2 A\right\}\chi_\Omega = 0 \quad \text{in } \mathbb{R}^2, \tag{2.5}$$

along with the "natural" boundary conditions

$$(\nabla - iA)\Psi\cdot\nu = 0 \quad \text{on } \partial\Omega, \tag{2.6}$$

$$\nu\times[\operatorname{curl} A] = 0 \quad \text{on } \partial\Omega. \tag{2.7}$$

Here $[\,\cdot\,]$ denotes the jump across $\partial\Omega$ and χ_Ω denotes the characteristic function of the set Ω. We note that the boundary condition (2.6) reduces to simply

$$\nabla\Psi\cdot\nu = 0 \quad \text{on } \partial\Omega \tag{2.8}$$

whenever the gauge choice (2.2) is used.

Of course, without further work, a critical point is only a weak solution of the system (2.4)–(2.7). However, standard elliptic regularity theory leads to the conclusion that any weak solution (Ψ, A) is in fact infinitely differentiable in the open sets Ω and $\mathbb{R}^2\setminus\overline{\Omega}$ and is a classical solution of (2.4)–(2.5) in these sets. The issue of boundary regularity is also not difficult but is perhaps a bit more subtle in light of the discontinuity inherent in (2.5) due to the presence of the characteristic function. As we shall use crucially the smoothness of both Ψ and A up to $\partial\Omega$, we present below the boundary regularity statement and proof.

PROPOSITION 2.1. *Assume $\Omega\subset\mathbb{R}^2$ is an open, bounded set with $\partial\Omega\in C^{5,\alpha}$ for some $\alpha>0$. Then any critical point (Ψ,A) of G in the sense of (2.3) expressed in the gauge (2.1) or (2.2) lies in the space $C^{3,\alpha}(\overline{\Omega};\mathbb{C})\times C^{1,\alpha}_{\text{loc}}(\mathbb{R}^2;\mathbb{R}^2)$. Furthermore, $A\in C^{3,\alpha}(\overline{\Omega};\mathbb{R}^2)$.*

Remark 2.1. We note that in general, A will not be smoother than $C^{1,\alpha}$ *across* $\partial\Omega$, so by claiming $A\in C^{3,\alpha}(\overline{\Omega};\mathbb{R}^2)$ we mean this in the sense of one-sided derivatives taken from within Ω.

Proof. Throughout this argument, when working with (2.4) for Ψ we will make the gauge choice (2.2), and when working with (2.5) for A we will choose (2.1). We note that switching gauges in no way affects the regularity of the solutions since the gauge transformations are smooth. In view of (2.1), one may rewrite (2.5) as

$$-\Delta A = f(x) \equiv \begin{cases} -\kappa^2 \operatorname{Im}(\Psi\nabla\Psi^*) - |\Psi|^2 A & \text{for } x\in\Omega, \\ 0 & \text{for } x\in\mathbb{R}^2\setminus\Omega. \end{cases} \tag{2.9}$$

Standard elliptic regularity immediately yields that $\Psi\in H^2(\Omega;\mathbb{C})$ while $A\in H^2_{\text{loc}}(\mathbb{R}^2;\mathbb{R}^2)$. Hence, in particular, $A\in C^{0,\alpha}_{\text{loc}}(\mathbb{R}^2;\mathbb{R}^2)$, meaning that Ψ satisfies a linear elliptic equation with Hölder continuous coefficients in Ω, along with homogeneous Neumann boundary conditions. Consequently, $\Psi\in C^{2,\alpha}(\overline{\Omega};\mathbb{C})$. This in turn implies that the function f defined by (2.9) lies in $L^\infty(\mathbb{R}^2;\mathbb{R}^2)$ so that $A\in C^{1,\alpha}_{\text{loc}}(\mathbb{R}^2;\mathbb{R}^2)$ (cf. [4, Chapters 6 and 8]).

Of course, both Ψ and A can be shown to be infinitely smooth in the open sets Ω and $\mathbb{R}^2\setminus\overline{\Omega}$, but our desire in this proposition is to establish smoothness up to $\partial\Omega$. Clearly the discontinuity in f across $\partial\Omega$ precludes higher regularity across the boundary so we restrict our attention to one-sided regularity from within Ω. To this end, fix any point $x_0 \in \partial\Omega$ and assume first that for some $R > 0$, the set $\partial\Omega \cap B(x_0, R)$ consists of the line segment

$$T = \{(x_1, x_2) : |x_1| < L,\ x_2 = 0\}$$

for some $L > 0$. Then locally, f takes the form

$$f(x_1, x_2) = \begin{cases} -\kappa^2 \mathrm{Im}\,(\Psi\nabla\Psi^*) - |\Psi|^2 A & \text{in } B(x_0, R) \cap \{x_2 > 0\}, \\ 0 & \text{in } B_(x_0, R) \cap \{x_2 < 0\}. \end{cases}$$

Choosing any test function $B \in H^1(\mathbb{R}^2; \mathbb{R}^2)$ supported in $B(x_0, R)$ we multiply (2.9) by $\frac{\partial B}{\partial x_1}$ and integrate by parts to find

$$\int_{B(x_0,R)} \nabla\left(\frac{\partial A}{\partial x_1}\right) \cdot \nabla B\, dx = \int_{B(x_0,R)} \frac{\partial f}{\partial x_1} \cdot B\, dx.$$

Here the crucial point is that f, while obviously not smooth, has an L^2—in fact, L^∞—derivative with respect to x_1. Consequently, $\frac{\partial A}{\partial x_1}$ is a weak solution to the equation

$$-\Delta\left(\frac{\partial A}{\partial x_1}\right) = \frac{\partial f}{\partial x_1} \in L^\infty(B(x_0, R)).$$

We conclude that $\frac{\partial A}{\partial x_1} \in C^{1,\alpha}(B(x_0, R); \mathbb{R}^2)$.

Now we shift our focus to the problem satisfied by A within Ω. Bringing to bear the above argument, we find that in $\Omega \cap B(x_0, R)$, A solves the problem (2.9) with $f \in C^{1,\alpha}$, along with a Dirichlet condition of class $C^{2,\alpha}(\partial\Omega \cap B(x_0, R); \mathbb{R}^2)$. As a result, we find $A \in C^{2,\alpha}(\overline{B(x_0, R) \cap \Omega}; \mathbb{R}^2)$.

Similarly, differentiating (2.4) along with the boundary condition (2.8) with respect to x_1, we conclude that $\Psi_{|\partial\Omega\cap B(x_0,R)}$ is of class $C^{3,\alpha}$ so that $\Psi \in C^{3,\alpha}(\overline{\Omega \cap B(x_0, R)}; \mathbb{C})$.

Differentiating (2.9) a second time with respect to x_1 and applying the same type of reasoning, we conclude that $A \in C^{3,\alpha}(\partial\Omega \cap B(x_0, R); \mathbb{R}^2)$; hence, $A \in C^{3,\alpha}(\overline{\Omega \cap B(x_0, R)}; \mathbb{R}^2)$.

As the argument is local, the case where $\partial\Omega$ is not locally flat is handled by a standard "flattening of the boundary" procedure. Keeping track of how many derivatives of the curvature κ are needed to carry this out, one finds that $C^{3,\alpha}$ regularity of the solution requires $\partial\Omega \in C^{5,\alpha}$ in order to apply the required Schauder theory. □

We conclude this section with some well-known facts about the Ginzburg–Landau system and a standard result relating the curvature of the boundary to functions satisfying a Neumann boundary condition.

Lemma 2.2. *Let $(\Psi, A) \in H^1(\Omega; \mathbb{C}) \times Z$ be a critical point of G. Then we have*

$$\text{(i)} \quad |\Psi| \le 1 \quad \text{in } \overline{\Omega}, \tag{2.10}$$

$$\text{(ii)} \quad \mathrm{curl}\, A = 0 \quad \text{in } \mathbb{R}^2 \setminus \overline{\Omega}. \tag{2.11}$$

The proof of (i) follows from the maximum principle after using (2.4) to obtain a differential inequality for $|\Psi|^2$ (cf. [3]). Property (ii) follows from the observation that in two dimensions, the condition $\mathrm{curl}\,\mathrm{curl}\, A = 0$ in $\mathbb{R}^2 \setminus \overline{\Omega}$ implies that the quantity $\mathrm{curl}\, A = (0, 0, \frac{\partial A^{(2)}}{\partial x_1} - \frac{\partial A^{(1)}}{\partial x_2})$ is constant. As $\mathrm{curl}\, A \in L^2(\mathbb{R}^2)$, the constant must be zero.

Lemma 2.3. *Let $B \in C^1(\overline{\Omega}; \mathbb{R}^2)$ satisfy the condition $B \cdot \nu = 0$ on $\partial\Omega$ and have zero curl on $\partial\Omega$, i.e., $\frac{\partial B^{(1)}}{\partial x_2} = \frac{\partial B^{(2)}}{\partial x_1}$ on $\partial\Omega$. Then we have the identity following on $\partial\Omega$:*

$$\frac{\partial\big(|B|^2\big)}{\partial \nu} = -2\frac{\partial \nu^{(k)}}{\partial x_j} B^{(k)} B^{(j)} = -2H\,|B|^2 . \tag{2.12}$$

In particular, if $u \in C^2(\overline{\Omega})$ satisfies the condition $\nabla u \cdot \nu = 0$ on $\partial\Omega$, then

$$\frac{\partial\big(|\nabla u|^2\big)}{\partial \nu} = -2\frac{\partial \nu^{(k)}}{\partial x_j} \frac{\partial u}{\partial x_k} \frac{\partial u}{\partial x_j} = -2H\,|\nabla u|^2 . \tag{2.13}$$

Proof. We compute

$$\frac{\partial\big(|B|^2\big)}{\partial \nu} = 2B^{(k)}\frac{\partial B^{(k)}}{\partial x_j}\nu^{(j)} = 2B^{(j)}\frac{\partial B^{(k)}}{\partial x_j}\nu^{(k)}. \tag{2.14}$$

Now $B \cdot \nu = 0$ implies

$$0 = \frac{\partial\big(B \cdot \nu\big)}{\partial x_j} B^{(j)} = B^{(j)}\frac{\partial B^{(k)}}{\partial x_j}\nu^{(k)} + B^{(k)}\frac{\partial \nu^{(k)}}{\partial x_j} B^{(j)},$$

and substituting this into (2.14) yields the result. □

3. Main results. In this section we will use the second variation of the Ginzburg–Landau energy to rule out the possibility of nontrivial stable critical points in convex planar domains. For any mappings $\Psi, \tilde{\Psi} \in H^1(\Omega; \mathbb{C})$ and $A, \tilde{A} \in Z$ (cf. 1.1) we denote by J the second variation of G taken about the pair (Ψ, A). We record here the straightforward calculation of J:

$$\begin{aligned} J(\Psi, A; \tilde{\Psi}, \tilde{A}) &= \frac{d^2}{d\varepsilon^2} G(\Psi + \varepsilon\tilde{\Psi}, A + \varepsilon\tilde{A})_{|\varepsilon=0} \\ &\frac{1}{2}\int_\Omega \bigg\{ \left|\nabla\tilde{\Psi}\right|^2 + i\langle \nabla\Psi, \tilde{\Psi}^*\tilde{A}\rangle + i\langle \nabla\tilde{\Psi}, A\tilde{\Psi}^* + \tilde{A}\Psi^*\rangle - i\langle \nabla\Psi^*, \tilde{A}\tilde{\Psi}\rangle \\ &\qquad - i\langle \nabla\tilde{\Psi}^*, A\tilde{\Psi} + \tilde{A}\Psi\rangle + |A|^2\left|\tilde{\Psi}\right|^2 + 2\langle A, \tilde{A}\rangle(\Psi\tilde{\Psi}^* + \Psi^*\tilde{\Psi}) + \left|\tilde{A}\right|^2|\Psi|^2 \bigg\}\,dx \\ &\qquad + \frac{\kappa^2}{4}\int_\Omega \bigg((\Psi\tilde{\Psi}^* + \tilde{\Psi}\Psi^*)^2 - 2(1 - |\Psi|^2)\left|\tilde{\Psi}\right|^2 \bigg)dx \\ &\qquad + \frac{1}{2}\int_{\mathbb{R}^2} \left|\operatorname{curl}\tilde{A}\right|^2\,dx. \end{aligned} \tag{3.1}$$

Throughout this section, we make the gauge choice (2.2).

Our approach hinges on a greatly simplified version of this formula when the second variation is taken about a critical point of the Ginzburg–Landau energy. This simplification will be accomplished in a few steps, the first of which is to reduce the expression to a sum of boundary integrals. As we will always be considering the second variation taken about a critical point (Ψ, A) of G, we will suppress this dependence and write J henceforth as a functional depending only on two arguments, i.e., we will simply write $J(\tilde{\Psi}, \tilde{A})$ rather than $J(\Psi, A, \tilde{\Psi}, \tilde{A})$.

Proposition 3.1. *Let (Ψ, A) be any critical point of G. Then we have*

$$J\left(\frac{\partial\Psi}{\partial x_j}, \frac{\partial A}{\partial x_j}\right) = \frac{1}{4}\int_{\partial\Omega} \frac{\partial}{\partial\nu}\left|\frac{\partial\Psi}{\partial x_j}\right|^2 ds + \frac{1}{2}\int_{\partial\Omega}\left\langle \nu \times \frac{\partial A}{\partial x_j}, \operatorname{curl}\left(\frac{\partial A}{\partial x_j}\right)\right\rangle ds$$
$$+ \frac{i}{4}\int_{\partial\Omega}\left\langle \frac{\partial A}{\partial x_j}, \nu\right\rangle \left(\Psi^*\frac{\partial\Psi}{\partial x_j} - \Psi\frac{\partial\Psi^*}{\partial x_j}\right) ds \quad (j = 1, 2). \tag{3.2}$$

Remark 3.1. In the second integral above, it is understood that $\operatorname{curl}(\frac{\partial A}{\partial x_j})$ denotes the trace from within Ω, which is well defined in light of the regularity theory developed in the previous section. A formula similar to (3.2) holds in three dimensions as well, provided that $\operatorname{curl}(\frac{\partial A}{\partial x_j}) \in L^2(\mathbb{R}^2)$, except that the quantity $\operatorname{curl}(\frac{\partial A}{\partial x_j})$ in the second boundary integral is replaced by the jump in this quantity across $\partial\Omega$.

Proof. Set $(\tilde\Psi, \tilde A) = (\partial\Psi/\partial x_j, \partial A/\partial x_j)$ in the second variation formula (3.1). Then we calculate

$$J(\partial\Psi/\partial x_j, \partial A/\partial x_j) = \frac{1}{2}\int_\Omega \left\{ \left|\nabla\left(\frac{\partial\Psi}{\partial x_j}\right)\right|^2 + i\left\langle \nabla\Psi, \frac{\partial\Psi^*}{\partial x_j}\frac{\partial A}{\partial x_j}\right\rangle \right.$$
$$+ i\left\langle \nabla\frac{\partial\Psi}{\partial x_j}, \frac{\partial\Psi^*}{\partial x_j}A + \Psi^*\frac{\partial A}{\partial x_j}\right\rangle - i\left\langle \nabla\Psi^*, \frac{\partial\Psi}{\partial x_j}\frac{\partial A}{\partial x_j}\right\rangle$$
$$- i\left\langle \nabla\frac{\partial\Psi^*}{\partial x_j}, \frac{\partial\Psi}{\partial x_j}A + \Psi\frac{\partial A}{\partial x_j}\right\rangle + |A|^2\left|\frac{\partial\Psi}{\partial x_j}\right|^2$$
$$\left. + 2\left\langle A, \frac{\partial A}{\partial x_j}\right\rangle\left(\Psi\frac{\partial\Psi^*}{\partial x_j} + \Psi^*\frac{\partial\Psi}{\partial x_j}\right) + |\Psi|^2\left|\frac{\partial A}{\partial x_j}\right|^2 \right\} dx$$
$$+ \frac{\kappa^2}{4}\int_\Omega \left(\left(\Psi\frac{\partial\Psi^*}{\partial x_j} + \Psi^*\frac{\partial\Psi}{\partial x_j}\right)^2 - 2(1 - |\Psi|^2)\left|\frac{\partial\Psi}{\partial x_j}\right|^2\right) dx$$
$$+ \frac{1}{2}\int_\Omega \left|\operatorname{curl}\left(\frac{\partial A}{\partial x_j}\right)\right|^2 dx.$$

Here we have used property (2.11) of Lemma 2.2 to conclude that $\operatorname{curl} A = 0$ in $\mathbb{R}^2\setminus\Omega$. We now denote the integrand above by $I_j(x)$. That is,

$$J\left(\frac{\partial\Psi}{\partial x_j}, \frac{\partial A}{\partial x_j}\right) = \int_\Omega I_j(x)\, dx. \tag{3.3}$$

Throughout this calculation, we shall invoke the regularity result, Proposition 2.1, which provides that both Ψ and A are C^3 up to the boundary. A straightforward calculation on $I_j(x)$ gives

$$I_j(x) = \frac{1}{4}\left(2\left|\nabla\frac{\partial\Psi}{\partial x_j}\right|^2 + i\left\langle \nabla\frac{\partial\Psi}{\partial x_j}, A\right\rangle\frac{\partial\Psi^*}{\partial x_j} - i\left\langle \nabla\frac{\partial\Psi^*}{\partial x_j}, A\right\rangle\frac{\partial\Psi}{\partial x_j} \right.$$
$$+ i\left\langle \nabla\Psi, \frac{\partial A}{\partial x_j}\right\rangle\frac{\partial\Psi^*}{\partial x_j} - i\left\langle \nabla\Psi^*, \frac{\partial A}{\partial x_j}\right\rangle\frac{\partial\Psi}{\partial x_j}$$
$$+ i\operatorname{div}\left(\Psi\frac{\partial A}{\partial x_j} + \frac{\partial\Psi}{\partial x_j}A\right)\frac{\partial\Psi^*}{\partial x_j} - i\operatorname{div}\left(\Psi^*\frac{\partial A}{\partial x_j} + \frac{\partial\Psi^*}{\partial x_j}A\right)\frac{\partial\Psi}{\partial x_j}$$
$$+ 2\left|\frac{\partial\Psi}{\partial x_j}\right|^2|A|^2 + 2\left\langle A, \frac{\partial A}{\partial x_j}\right\rangle\left(\Psi\frac{\partial\Psi^*}{\partial x_j} + \Psi^*\frac{\partial\Psi}{\partial x_j}\right)$$

$$
\begin{aligned}
&+\kappa^2\left(\left(\Psi\frac{\partial\Psi^*}{\partial x_j}+\Psi^*\frac{\partial\Psi}{\partial x_j}\right)^2-2\left(1-|\Psi|^2\right)\left|\frac{\partial\Psi}{\partial x_j}\right|^2\right)\Bigg)\\
&+\frac{1}{2}\left(\left|\operatorname{curl}\left(\frac{\partial A}{\partial x_j}\right)\right|^2+|\Psi|^2\left|\frac{\partial A}{\partial x_j}\right|^2+\left\langle A,\frac{\partial A}{\partial x_j}\right\rangle\left(\Psi\frac{\partial\Psi^*}{\partial x_j}+\Psi^*\frac{\partial\Psi}{\partial x_j}\right)\right.\\
&+\frac{i}{4}\left(\left\langle\nabla\Psi,\frac{\partial A}{\partial x_j}\right\rangle\frac{\partial\Psi^*}{\partial x_j}+\left\langle\nabla\frac{\partial\Psi}{\partial x_j},\frac{\partial A}{\partial x_j}\right\rangle\Psi^*\right.\\
&\left.\left.-\left\langle\nabla\Psi^*,\frac{\partial A}{\partial x_j}\right\rangle\frac{\partial\Psi}{\partial x_j}-\left\langle\nabla\frac{\partial\Psi^*}{\partial x_j},\frac{\partial A}{\partial x_j}\right\rangle\Psi\right)\right)\\
&+\tilde{I}_j(x),
\end{aligned}
\tag{3.4}
$$

where

$$
\begin{aligned}
\tilde{I}_j(x)=\frac{i}{4}&\left(\left\langle\nabla\frac{\partial\Psi}{\partial x_j},\frac{\partial A}{\partial x_j}\right\rangle\Psi^*-\left\langle\nabla\frac{\partial\Psi^*}{\partial x_j},\frac{\partial A}{\partial x_j}\right\rangle\Psi-\left\langle\nabla\Psi,\frac{\partial A}{\partial x_j}\right\rangle\frac{\partial\Psi^*}{\partial x_j}\right.\\
&\left.+\left\langle\nabla\Psi^*,\frac{\partial A}{\partial x_j}\right\rangle\frac{\partial\Psi}{\partial x_j}\right)-\frac{i}{4}\operatorname{div}\left(\frac{\partial A}{\partial x_j}\right)\left(\Psi\frac{\partial\Psi^*}{\partial x_j}-\Psi^*\frac{\partial\Psi}{\partial x_j}\right).
\end{aligned}
$$

The expression $\tilde{I}_j(x)$ is easily simplified to

$$
\tilde{I}_j(x)=-\frac{i}{4}\operatorname{div}\left(\left(\Psi\frac{\partial\Psi^*}{\partial x_j}-\Psi^*\frac{\partial\Psi}{\partial x_j}\right)\frac{\partial A}{\partial x_j}\right).
\tag{3.5}
$$

On the other hand, by differentiating the Ginzburg–Landau system with respect to x_j, we have the following identities:

$$
\begin{aligned}
\Delta\left(\frac{\partial\Psi}{\partial x_j}\right)&-i\left\langle\nabla\left(\frac{\partial\Psi}{\partial x_j}\right),A\right\rangle-i\left\langle\nabla\Psi,\frac{\partial A}{\partial x_j}\right\rangle-i\operatorname{div}\left(\frac{\partial A}{\partial x_j}\Psi+A\frac{\partial\Psi}{\partial x_j}\right)\\
&-2\left\langle A,\frac{\partial A}{\partial x_j}\right\rangle\Psi-|A|^2\frac{\partial\Psi}{\partial x_j}+\kappa^2\left(-\frac{\partial\Psi}{\partial x_j}\Psi^*-\frac{\partial\Psi^*}{\partial x_j}\Psi\right)\Psi\\
&+\kappa^2(1-|\Psi|^2)\frac{\partial\Psi}{\partial x_j}\\
&=0\quad\text{in }\Omega,\\
\operatorname{curl}\operatorname{curl}\left(\frac{\partial A}{\partial x_j}\right)&+|\Psi|^2\frac{\partial A}{\partial x_j}+\left(\frac{\partial\Psi}{\partial x_j}\Psi^*+\Psi\frac{\partial\Psi^*}{\partial x_j}\right)A\\
&+\frac{i}{2}\left(\frac{\partial\Psi^*}{\partial x_j}\nabla\Psi+\Psi^*\nabla\frac{\partial\Psi}{\partial x_j}-\frac{\partial\Psi}{\partial x_j}\nabla\Psi^*-\Psi\nabla\frac{\partial\Psi^*}{\partial x_j}\right)\\
&=0\quad\text{in }\Omega.
\end{aligned}
$$

These lead readily to the relations

$$
\begin{aligned}
-\Delta\left(\frac{\partial\Psi}{\partial x_j}\right)\frac{\partial\Psi^*}{\partial x_j}&-\Delta\left(\frac{\partial\Psi^*}{\partial x_j}\right)\frac{\partial\Psi}{\partial x_j}+i\left\langle\nabla\frac{\partial\Psi}{\partial x_j},A\right\rangle\frac{\partial\Psi^*}{\partial x_j}+i\left\langle\nabla\Psi,\frac{\partial A}{\partial x_j}\right\rangle\frac{\partial\Psi^*}{\partial x_j}\\
&-i\left\langle\nabla\frac{\partial\Psi^*}{\partial x_j},A\right\rangle\frac{\partial\Psi}{\partial x_j}-i\left\langle\nabla\Psi^*,\frac{\partial A}{\partial x_j}\right\rangle\frac{\partial\Psi}{\partial x_j}\\
&+i\operatorname{div}\left(\Psi\frac{\partial A}{\partial x_j}+\frac{\partial\Psi}{\partial x_j}A\right)\frac{\partial\Psi^*}{\partial x_j}-i\operatorname{div}\left(\Psi^*\frac{\partial A}{\partial x_j}+\frac{\partial\Psi^*}{\partial x_j}A\right)\frac{\partial\Psi}{\partial x_j}
\end{aligned}
$$

$$\begin{aligned}
&+2\left|\frac{\partial\Psi}{\partial x_j}\right|^2|A|^2+2\left\langle A,\frac{\partial A}{\partial x_j}\right\rangle\left(\Psi\frac{\partial\Psi^*}{\partial x_j}+\Psi^*\frac{\partial\Psi}{\partial x_j}\right)\\
&+\kappa^2\left(\left(\Psi\frac{\partial\Psi^*}{\partial x_j}+\Psi^*\frac{\partial\Psi}{\partial x_j}\right)^2-2\left(1-|\Psi|^2\right)\left|\frac{\partial\Psi}{\partial x_j}\right|^2\right)\\
&=0\quad\text{in }\Omega,
\end{aligned}\tag{3.6}$$

$$\begin{aligned}
\operatorname{curl}\operatorname{curl}\left(\frac{\partial A}{\partial x_j}\right)\frac{\partial A}{\partial x_j}&+|\Psi|^2\left|\frac{\partial A}{\partial x_j}\right|^2+\left\langle A,\frac{\partial A}{\partial x_j}\right\rangle\left(\Psi\frac{\partial\Psi^*}{\partial x_j}+\Psi^*\frac{\partial\Psi}{\partial x_j}\right)\\
&+\frac{i}{2}\left(\left\langle\nabla\Psi,\frac{\partial A}{\partial x_j}\right\rangle\frac{\partial\Psi^*}{\partial x_j}+\left\langle\nabla\frac{\partial\Psi}{\partial x_j},\frac{\partial A}{\partial x_j}\right\rangle\Psi^*\right.\\
&\qquad\left.-\left\langle\nabla\Psi^*,\frac{\partial A}{\partial x_j}\right\rangle\frac{\partial\Psi}{\partial x_j}-\left\langle\nabla\frac{\partial\Psi^*}{\partial x_j},\frac{\partial A}{\partial x_j}\right\rangle\Psi\right)\\
&=\mathbf{0}\quad\text{in }\Omega.
\end{aligned}\tag{3.7}$$

We shall also need the following identities, which are an immediate consequence of integration by parts:

$$\begin{aligned}
\int_\Omega 2\left|\frac{\partial\Psi}{\partial x_j}\right|^2dx&=\int_{\partial\Omega}\left(\frac{\partial\Psi}{\partial x_j}\frac{\partial}{\partial\nu}\left(\frac{\partial\Psi^*}{\partial x_j}\right)+\frac{\partial\Psi^*}{\partial x_j}\frac{\partial}{\partial\nu}\left(\frac{\partial\Psi^*}{\partial x_j}\right)\right)ds\\
&\quad-\int_\Omega\left(\frac{\partial\Psi}{\partial x_j}\Delta\left(\frac{\partial\Psi^*}{\partial x_j}\right)+\frac{\partial\Psi^*}{\partial x_j}\Delta\left(\frac{\partial\Psi}{\partial x_j}\right)\right)dx\\
&=\int_{\partial\Omega}\frac{\partial}{\partial\nu}\left|\frac{\partial\Psi}{\partial x_j}\right|^2ds-\int_\Omega\left(\frac{\partial\Psi}{\partial x_j}\Delta\left(\frac{\partial\Psi^*}{\partial x_j}\right)+\frac{\partial\Psi^*}{\partial x_j}\Delta\left(\frac{\partial\Psi}{\partial x_j}\right)\right)dx,
\end{aligned}\tag{3.8}$$

$$\begin{aligned}
\int_\Omega\left|\operatorname{curl}\frac{\partial A}{\partial x_j}\right|^2dx&=\int_{\partial\Omega}\left\langle\nu\times\frac{\partial A}{\partial x_j},\operatorname{curl}\left(\frac{\partial A}{\partial x_j}\right)\right\rangle ds\\
&\quad+\int_\Omega\left\langle\frac{\partial A}{\partial x_j},\operatorname{curl}\operatorname{curl}\left(\frac{\partial A}{\partial x_j}\right)\right\rangle dx.
\end{aligned}\tag{3.9}$$

If we use the identities (3.6), (3.7), (3.8), and (3.9) in (3.3) and (3.4), we arrive at a simple expression:

$$J\left(\frac{\partial\Psi}{\partial x_j},\frac{\partial A}{\partial x_j}\right)=\frac{1}{4}\int_{\partial\Omega}\frac{\partial}{\partial\nu}\left|\frac{\partial\Psi}{\partial x_j}\right|^2ds+\frac{1}{2}\int_{\partial\Omega}\left\langle\nu\times\frac{\partial A}{\partial x_j},\operatorname{curl}\left(\frac{\partial A}{\partial x_j}\right)\right\rangle ds+\int_\Omega\tilde{I}_j(x)\,dx.$$

Then applying the divergence theorem to (3.5), we see that the last term of this expression is equal to

$$-\frac{i}{4}\int_{\partial\Omega}\left\langle\frac{\partial A}{\partial x_j},\nu\right\rangle\left(\Psi\frac{\partial\Psi^*}{\partial x_j}-\Psi^*\frac{\partial\Psi}{\partial x_j}\right)ds.$$

This completes the proof of the proposition. □

In the next step, we will invoke the assumption of convexity of $\partial\Omega$ to establish nonpositivity of the second variation.

PROPOSITION 3.2. *Let (Ψ,A) be any critical point of G. Denote $\Gamma_1=\{x\in\partial\Omega:\Psi(x)\neq 0\}$ and $\Gamma_2=\{x\in\partial\Omega:\Psi(x)=0\}$. Also, on Γ_1 write Ψ locally as*

$\Psi(x) = we^{i\phi}$. Then we have

$$\sum_{j=1}^{2} J(\partial\Psi/\partial x_j, \partial A/\partial x_j) = -\frac{1}{2}\int_{\Gamma_1} H\Big\{|\nabla w|^2 + w^2\,|\nabla\phi - A|^2\Big\}\,ds - \frac{1}{2}\int_{\Gamma_2} H\,|\nabla\Psi|^2\,ds, \tag{3.10}$$

where $H = H(x)$ denotes the curvature of $\partial\Omega$. In particular, for a convex domain Ω whereby $H \geq 0$, we have

$$\sum_{j=1}^{2} J(\partial\Psi/\partial x_j, \partial A/\partial x_j) \leq 0. \tag{3.11}$$

Remark 3.2. A similar formula holds in three dimensions as well, but it contains an extra term related to the fact that $\operatorname{curl} A$ does not necessarily vanish outside $\Omega \subset \mathbb{R}^3$.

Proof. We make the gauge choice (2.2). Then taking formula (3.2) of Proposition 3.1 as a starting point, we let $K \equiv \sum_{j=1}^{2}\langle \nu \times \frac{\partial A}{\partial x_j}, \operatorname{curl}(\frac{\partial A}{\partial x_j})\rangle$ and use the conditions $\operatorname{div} \mathrm{A} = 0$ and $\nabla\operatorname{div} \mathrm{A} = 0$ in $\overline{\Omega}$ to calculate

$$\begin{aligned} K &= \left(\nu_1 \frac{\partial A^{(2)}}{\partial x_1} - \nu_2 \frac{\partial A^{(1)}}{\partial x_1}\right)\left(\frac{\partial^2 A^{(2)}}{\partial x_1^2} - \frac{\partial^2 A^{(1)}}{\partial x_1 \partial x_2}\right) \\ &\quad + \left(\nu_1 \frac{\partial A^{(2)}}{\partial x_2} - \nu_2 \frac{\partial A^{(1)}}{\partial x_2}\right)\left(\frac{\partial^2 A^{(2)}}{\partial x_1 \partial x_2} - \frac{\partial^2 A^{(1)}}{\partial x_2^2}\right) \\ &= \left(\nu_1 \frac{\partial A^{(2)}}{\partial x_1} + \nu_2 \frac{\partial A^{(2)}}{\partial x_2}\right)\left(\frac{\partial^2 A^{(2)}}{\partial x_1^2} + \frac{\partial^2 A^{(2)}}{\partial x_2^2}\right) \\ &\quad + \left(-\nu_1 \frac{\partial A^{(1)}}{\partial x_1} - \nu_2 \frac{\partial A^{(1)}}{\partial x_2}\right)\left(-\frac{\partial^2 A^{(1)}}{\partial x_1^2} - \frac{\partial^2 A^{(1)}}{\partial x_2^2}\right). \end{aligned}$$

Hence, we get

$$K = \sum_{\ell=1}^{2} \frac{\partial A^{(\ell)}}{\partial \nu}\Delta A^{(\ell)}.$$

On the other hand, our gauge choice also simplifies the Ginzburg–Landau equation (2.5) to read

$$\Delta A = |\Psi|^2 A + \frac{i}{2}(\Psi^*\nabla\Psi - \Psi\nabla\Psi^*) \quad \text{in } \Omega,$$

and substituting this into K, we get

$$K = \sum_{\ell=1}^{2} \frac{\partial A^{(\ell)}}{\partial \nu}\left(|\Psi|^2 A^{(\ell)} + \frac{i}{2}\left(\Psi^*\frac{\partial\Psi}{\partial x_\ell} - \Psi\frac{\partial\Psi^*}{\partial x_\ell}\right)\right) \quad \text{on } \partial\Omega.$$

Then from Proposition 3.1 we get

$$\begin{aligned} \sum_{j=1}^{2} J\left(\frac{\partial\Psi}{\partial x_j}, \frac{\partial A}{\partial x_j}\right) &= \frac{1}{4}\int_{\partial\Omega} \frac{\partial}{\partial\nu}|\nabla\Psi|^2\,ds + \frac{1}{4}\int_{\partial\Omega}|\Psi|^2\frac{\partial}{\partial\nu}|A|^2\,ds \\ &\quad + \frac{i}{4}\int_{\partial\Omega}\sum_{j=1}^{2}\left(\frac{\partial A^{(j)}}{\partial\nu} + \left\langle\frac{\partial A}{\partial x_j}, \nu\right\rangle\right)\left(\Psi^*\frac{\partial\Psi}{\partial x_j} - \Psi\frac{\partial\Psi^*}{\partial x_j}\right)ds. \end{aligned} \tag{3.12}$$

Now define

$$F(x) = \frac{\partial}{\partial \nu}|\nabla\Psi|^2 + |\Psi|^2\frac{\partial}{\partial \nu}|A|^2 + i\sum_{j=1}^{2}\left(\frac{\partial A^{(j)}}{\partial \nu} + \left\langle \frac{\partial A}{\partial x_j}, \nu\right\rangle\right)\left(\Psi^*\frac{\partial \Psi}{\partial x_j} - \Psi\frac{\partial \Psi^*}{\partial x_j}\right)$$

on $\partial\Omega$. By using the condition $\operatorname{curl} A = 0$ on $\partial\Omega$, we have

$$\frac{\partial A^{(j)}}{\partial \nu} = \frac{\partial A^{(j)}}{\partial x_\ell}\nu^{(\ell)} = \frac{\partial A^{(\ell)}}{\partial x_j}\nu^{(\ell)} = \left\langle \frac{\partial A}{\partial x_j}, \nu\right\rangle.$$

Hence F can also be written as

$$F(x) = \frac{\partial}{\partial \nu}|\nabla\Psi|^2 + |\Psi|^2\frac{\partial}{\partial \nu}|A|^2 + 2i\sum_{j=1}^{2}\left\langle \frac{\partial A}{\partial x_j}, \nu\right\rangle\left(\Psi^*\frac{\partial \Psi}{\partial x_j} - \Psi\frac{\partial \Psi^*}{\partial x_j}\right). \tag{3.13}$$

We calculate $F(x)$ separately on Γ_1 and Γ_2, where

$$\text{(i)}\quad \Gamma_1 = \{x \in \partial\Omega \mid \Psi(x) \neq 0\},$$
$$\text{(ii)}\quad \Gamma_2 = \{x \in \partial\Omega \mid \Psi(x) = 0\}.$$

First consider any point $x_0 \in \Gamma_1$ and consider a small contractible neighborhood V of x_0 so that Ψ does not vanish in $V \cap \Omega$. In this situation we can write $\Psi(x)$ in the form $\Psi(x) = w(x)e^{i\,\phi(x)}$ and we see that

(3.14)
$$|\nabla\Psi|^2 = |\nabla w(x)|^2 + w(x)^2|\nabla\phi|^2, \quad \Psi^*\frac{\partial \Psi}{\partial x_j} - \Psi\frac{\partial \Psi^*}{\partial x_j} = 2iw(x)^2\frac{\partial \phi}{\partial x_j} \quad \text{in} \quad V \cap \Omega.$$

We also note that ϕ and w satisfy the Neumann boundary condition on $\Gamma_1 \cap V$ and we recall that $A \cdot \nu = 0$ on $\partial\Omega$ as well. Hence, we find

$$0 = \nabla(A\cdot\nu)\cdot\nabla\phi = \frac{\partial A^{(k)}}{\partial x_j}\nu^{(k)}\frac{\partial \phi}{\partial x_j} + A^{(k)}\frac{\partial \nu^{(k)}}{\partial x_j}\frac{\partial \phi}{\partial x_j}. \tag{3.15}$$

In order to evaluate the expression $F(x_0)$ we assume, without loss of generality, that $\nu(x_0) = (0, 1)$. In particular, this implies that

$$\frac{\partial w}{\partial x_2}(x_0) = \frac{\partial \phi}{\partial x_2}(x_0) = A^{(2)}(x_0) = 0 \text{ and that } \frac{\partial \nu^{(1)}}{\partial x_1}(x_0) = H. \tag{3.16}$$

Then, by Lemma 2.3, (3.14), (3.15), and (3.16), we get

$$\begin{aligned}
F(x_0) &= \frac{\partial}{\partial \nu}(|\nabla w|^2 + w^2|\nabla\phi|^2) + w^2\frac{\partial}{\partial \nu}|A|^2 - 4w^2\sum_{j=1}^{2}\frac{\partial \phi}{\partial x_j}\left\langle \frac{\partial A}{\partial x_j}, \nu\right\rangle \\
&= -2\frac{\partial \nu^{(k)}}{\partial x_j}\frac{\partial w}{\partial x_k}\frac{\partial w}{\partial x_j} - 2w^2\frac{\partial \nu^{(k)}}{x_j}\left(\frac{\partial \phi}{\partial x_k}\frac{\partial \phi}{\partial x_j} + A^{(k)}A^{(j)} - 2A^{(k)}\frac{\partial \phi}{\partial x_j}\right) \\
&= -2\frac{\partial \nu^{(1)}}{\partial x_1}\left(\left(\frac{\partial w}{\partial x_1}\right)^2 + w^2\left(\frac{\partial \phi}{\partial x_1} - A^{(1)}\right)^2\right) \\
&= -2H(|\nabla w|^2 + w^2\,|\nabla\phi - A|^2) \quad \text{in } V \cap \Gamma_1.
\end{aligned} \tag{3.17}$$

Now consider $x_0 \in \Gamma_2$ where then $\Psi(x_0) = 0$. From Lemma 2.3 and (3.13) we get

$$\begin{aligned} F(x_0) &= \frac{\partial}{\partial\nu}|\nabla\Psi|^2 = \frac{\partial}{\partial\nu}|\nabla(\mathrm{Re}\Psi)|^2 + \frac{\partial}{\partial\nu}|\nabla(\mathrm{Im}\Psi)|^2 \\ &= -2H\,|\nabla(\mathrm{Re}\Psi)|^2 - 2H\,|\nabla(\mathrm{Im}\Psi)|^2 = -2H\,|\nabla\Psi|^2 \quad \text{on } \Gamma_2. \end{aligned} \tag{3.18}$$

The formula (3.10) follows from the substitution of (3.17) and (3.18) into (3.12). □

We now restate and prove our main result.

THEOREM 3.3. *Let $\Omega \subset \mathbb{R}^2$ be a bounded, open, convex set with $\partial\Omega \in C^{5,\alpha}$ for some $\alpha \in (0,1)$. Then the only critical point (Ψ, A) of the Ginzburg–Landau energy G for which the second variation J is nonnegative is the trivial one $(1,0)$ and its gauge-equivalent representations.*

Proof. We fix the gauge choice (2.2) and let (Ψ, A) be any critical point for which the second variation is nonnegative. Then in light of Proposition 3.2, we can conclude that

$$J(\partial\Psi/\partial x_j, \partial A/\partial x_j) = 0 \quad \text{for } j = 1,2, \tag{3.19}$$

so that in fact $(\partial\Psi/\partial x_j, \partial A/\partial x_j)$ is a minimizer of the second variation (3.1) for both $j = 1$ and $j = 2$. In particular, this implies that $(\partial\Psi/\partial x_j, \partial A/\partial x_j)$ satisfies the natural boundary conditions associated with critical points of (3.1), namely,

$$\left(\nabla\left(\frac{\partial\Psi}{\partial x_j}\right) - i\Psi\frac{\partial A}{\partial x_j}\right)\cdot\nu = 0 \quad \text{on } \partial\Omega \text{ for } j = 1,2. \tag{3.20}$$

Let us denote by Γ_3 the nonempty, relatively open subset of $\partial\Omega$ given by

$$\Gamma_3 = \{x \in \partial\Omega : H(x) > 0\}.$$

It follows from (3.10) and (3.19) that

$$\nabla\Psi(x) - iA\Psi(x) = 0 \quad \text{for } x \in \Gamma_3. \tag{3.21}$$

Now let us decompose Γ_3 into a union $\Gamma_4 \cup \Gamma_5$ where

$$\Gamma_4 = \Gamma_3 \cap \{x \in \partial\Omega : \Psi \neq 0\} \quad \text{and} \quad \Gamma_5 = \Gamma_3 \cap \{x \in \partial\Omega : \Psi = 0\}.$$

We first pursue the possibility that $\Gamma_4 \neq \emptyset$ and consider a point $x_0 \in \Gamma_4$. For some $\varepsilon > 0$ we may express Ψ in $\overline{\Omega} \cap B(x_0, \varepsilon)$ as $\Psi = we^{i\phi}$ and from (2.10) and (3.21) it follows that $w \equiv c$ on $\partial\Omega \cap B(x_0,\varepsilon)$ for some $c \in (0,1]$.

From (3.20) one concludes that

$$\frac{\partial^2 w}{\partial x_j \partial x_k}\nu^{(k)} = 0 \quad \text{on } \partial\Omega \cap B(x_0,\varepsilon) \text{ for } j = 1,2, \tag{3.22}$$

while from tangential differentiation of (3.21) one finds that

$$\frac{\partial^2 w}{\partial x_j \partial x_k}\tau^{(k)} = 0 \quad \text{on } \partial\Omega \cap B(x_0,\varepsilon) \text{ for } j = 1,2, \tag{3.23}$$

where we have let τ denote the unit tangent vector to $\partial\Omega$ so that $(\tau^{(1)}, \tau^{(2)}) = (-\nu^{(2)}, \nu^{(1)})$. Together, (3.22) and (3.23) imply that all second partials of w vanish on $\partial\Omega \cap B(x_0, \varepsilon)$.

Now on the set $\Omega \cap B(x_0, \varepsilon)$ one finds from (2.4) and (2.6) that

$$\Delta w - |\nabla\phi - A|^2 w + \kappa^2 w(1 - w^2) = 0 \quad \text{in } \Omega \cap B(x_0, \varepsilon), \tag{3.24}$$

$$\partial w/\partial\nu = 0 \quad \text{on } \partial\Omega \cap B(x_0, \varepsilon). \tag{3.25}$$

It then follows immediately from (3.21), (3.24), and the fact that Ψ is C^2 up to the boundary that $w \equiv 1$ on $\Gamma_4 \cap B(x_0, \varepsilon)$. Viewing (3.24) as an elliptic equation of the form

$$\Delta(1 - w) + q(x)(1 - w) \leq 0 \quad \text{in } B(x_0, \varepsilon) \cap \Omega$$

with $q(x) = -\kappa^2 w(1+w)$ one can apply the strong maximum principle to the nonnegative function $1 - w$ to conclude that either $w \equiv 1$ in $B(x_0, \varepsilon) \cap \Omega$ or else $1 - w > 0$ in $B(x_0, \varepsilon) \cap \Omega$. The latter possibility is then eliminated using the Hopf maximum principle, in light of the boundary condition (3.25) (cf. [4, Lemma 3.4 and Theorem 3.5]). Hence, $w \equiv 1$ in $B(x_0, \varepsilon) \cap \overline{\Omega}$ and by (3.24) it follows that $|\nabla\phi - A| \equiv 0$ on this set as well. By continuation we can then extend these relations throughout Ω because it is simply connected. Thus, $(\Psi, A) = (e^{i\phi}, \nabla\phi)$ throughout Ω for some real-valued function ϕ, which is our desired conclusion.

Finally, we consider the possibility that $\Gamma_4 = \emptyset$. Then $\Gamma_5 = \Gamma_3$. Writing $\Psi = u + iv$, we find that the real functions u and v satisfy the system

$$\begin{aligned} &\Delta u + 2A \cdot \nabla v - |A|^2 u + \kappa^2(1 - (u^2 + v^2))u = 0, \\ &\Delta v - 2A \cdot \nabla u - |A|^2 v + \kappa^2(1 - (u^2 + v^2))v = 0. \end{aligned} \tag{3.26}$$

Furthermore, fixing $x_0 \in \Gamma_5$ and a sufficiently small $\varepsilon > 0$ one finds that u and v satisfy the boundary conditions

$$u = v = \nabla u \cdot \nu = \nabla v \cdot \nu = 0 \quad \text{on } \partial\Omega \cap B(x_0, \varepsilon).$$

Invoking Calderón's uniqueness theorem (cf. [1]), we conclude that $u = v = 0$ in a neighborhood of x_0. Again, in light of the simple-connectivity of Ω, we conclude that $u = v = 0$ throughout Ω. Turning to (2.5), we conclude that $\operatorname{curl}\operatorname{curl} A = 0$ throughout $\mathbb{R}^2$. As in Lemma 2.2, we find that $\operatorname{curl} A = 0$ in $\mathbb{R}^2$ and so $A = \nabla\eta$ for some function $\eta : \mathbb{R}^2 \to \mathbb{R}^1$. Hence (Ψ, A) is gauge-equivalent to $(0, 0)$. But $(0, 0)$ is clearly unstable in view of (3.1) since

$$J(0, 0; 1, 0) = -\frac{\kappa^2}{2} |\Omega|.$$

Thus, the only possibility for a stable critical point in a convex domain is that (Ψ, A) is gauge-equivalent to $(1, 0)$. □

Remark 3.3. We suspect that the same nonexistence result holds in three-dimensional convex domains and we are presently investigating this possibility. See Remark (3.2).

REFERENCES

[1] A. Calderón, *Uniqueness in the Cauchy problem for partial differential equations*, Amer. J. Math., 80 (1958), pp. 16–36.

[2] R. Casten and C. Holland, *Instability results for reaction-diffusion equations with Neumann boundary conditions*, J. Differential Equations, 27 (1978), pp. 266–273.

[3] 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.

[4] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, New York, 1983.

[5] V. L. Ginzburg and L. D. Landau, *On the theory of superconductivity*, J.E.T.P., 20 (1950), p. 1064.

[6] S. Jimbo and Y. Morita, *Ginzburg–Landau equations and stable solutions in a rotational domain*, SIAM J. Math. Anal., 27 (1996), pp. 1360–1385.

[7] S. Jimbo and J. Zhai, *Ginzburg-Landau equation with magnetic effect: Non-simply-connected domains*, J. Math. Soc. Japan, 50 (1998), pp. 663–684.

[8] S. Jimbo and J. Zhai, *Domain perturbation method and local minimizers to Ginzburg-Landau functional with magnetic effect*, Abstr. Appl. Anal., 5 (2000), pp. 101–112.

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

[10] J. Rubinstein and P. Sternberg, *Homotopy classification of minimizers of the Ginzburg-Landau energy and the existence of permanent currents*, Comm. Math. Phys., 179 (1996), pp. 257–263.

SIAM J. Math. Anal.
Vol. 33, No. 6, pp. 1393–1410

SCATTERED DATA INTERPOLATION ON SPHERES: ERROR ESTIMATES AND LOCALLY SUPPORTED BASIS FUNCTIONS[*]

FRANCIS J. NARCOWICH[†] AND JOSEPH D. WARD[†]

Abstract. Error estimates for scattered data interpolation by "shifts" of a positive definite function for target functions in the associated reproducing kernel Hilbert space (RKHS) have been known for a long time. However, apart from special cases where data is gridded, these estimates do not apply when the target functions generating the data are outside of the associated RKHS, and in fact no estimates were known for such target functions. In this paper, working with the n-sphere as the underlying manifold, we obtain Sobolev-type error estimates for interpolating functions $f \in C^{2k}(S^n)$ from "shifts" of a smoother positive definite function ϕ defined on S^n. Moreover, the estimates are close to the optimal approximation order. We also introduce a class of locally supported positive definite functions on S^n, functions based on Wendland's compactly supported radial basis functions (RBFs) [H. Wendland, *Adv. Comput. Math.*, 4 (1995), pp. 389–396], which can be both explicitly and easily computed and also analyzed for convergence properties.

Key words. interpolation on spheres, error estimates, rates of convergence, locally supported basis functions, scattered data

AMS subject classifications. 41A25, 41A05, 41A63, 42C10

PII. S0036141001395054

1. Introduction.

The problem of effectively representing an underlying function based on its sampled values is both old and important. One way to do this is to interpolate the data from a class of functions. The success of this approach is based on many criteria including the cost of producing the interpolant, the robustness of the interpolation process, and how well the interpolant approximates the underlying function. These are the issues we deal with in this paper when the domain of the underlying function is S^n.

Radial basis functions (RBFs) have proven to be a powerful tool for analyzing scattered data on R^n. More recently, spherical basis functions (SBFs), which are analogs of RBFs on the n-sphere, and periodic basis functions (PBFs), which are analogs of RBFs on the n-torus, have had comparable success for analyzing scattered data on these manifolds. In all cases, interpolants by "shifts" of an RBF, SBF, or PBF are constructed from sampling a function f at scattered sites. If f belongs to a certain reproducing kernel Hilbert space (RKHS) associated with ϕ, N_ϕ, the *native space* of ϕ, then these interpolants will converge to f. However, if ϕ is smooth, the RKHS N_ϕ is small in the sense that it is composed of very smooth functions. Up to

[*] Received by the editors September 12, 2001; accepted for publication (in revised form) February 1, 2002; published electronically July 9, 2002. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research or the U.S. Government.

http://www.siam.org/journals/sima/33-6/39505.html

[†] Department of Mathematics, Texas A&M University, College Station, TX 77843-3368 (fnarc@math.tamu.edu, jward@math.tamu.edu). The first author's research was sponsored by the Air Force Office of Scientific Research, Air Force Office of Material Command, USAF, under grant F49620-98-1-0204. The second author's research was supported by grant DMS-9971276 from the National Science Foundation.

now, any error analysis involving scattered data has thus been limited to such smooth functions.

One of the purposes of this paper is to remove in the case of S^n the "native space barrier" for a very large class of SBFs. When an SBF ϕ has Legendre coefficients that are comparable to reciprocals of Sobolev weights of some order τ, we obtain, in section 3, interpolation errors for functions lying outside the native space: Suppose that $f \in C^{2k}(S^n)$ is sampled on a finite number of distinct points comprising a set $X \subset S^n$. If any point in S^n is at most a distance h from X and if the distance between points in X is comparable to h, then the interpolant $I_X f$ satisfies

$$\|f - I_X f\| \leq \text{const} \cdot h^{2k-n/2} \|\Delta^k f\|$$

for $\tau \geq 2k > \frac{n}{2}$, where here and elsewhere in this paper $\|\cdot\| := \|\cdot\|_{C(S^n)}$. (See Theorem 3.2 for the precise statement.) The key to this approach is a result on simultaneous approximation and interpolation by spherical harmonics; this and the estimates themselves are discussed in section 3. We also make the point that for the class N_ϕ, $h^{\tau-n/2}$ is optimal. In this respect, our results resemble ones that hold for splines.

While our results do not yet apply to the R^n case, it seems very likely that a variant of this approach should be applicable there as well. The authors are currently working in this direction. In this connection, we mention a paper of Yoon [27], which was pointed out to us by one of the referees. There, Yoon investigates error estimates in R^n for a special class of radial functions, namely thin-plate splines that depend on a parameter λ. His focus is different from ours. He works on cases in which the parameter λ is required to depend on the spacing of the data. In effect, the radial function is changing with the data.

Another purpose of this paper is to introduce a class of locally supported SBFs that can be both explicitly computed and analyzed. It is easy to produce a large class of SBFs on S^n that are explicit (thus easy to evaluate) by restricting a given RBF on R^{n+1} to S^n by means of

$$\phi(x \cdot y) = \Phi(\|x - y\|_2)|_{x,y \in S^n}.$$

On the other hand, the basic tools for analyzing rates of convergence of interpolants and stability of interpolation matrices are all based on the rates of decay of the Legendre coefficients for $\phi(x \cdot y) = \phi(\cos\theta)$. It is easy to construct SBFs that are defined by their Legendre series and that satisfy various decay rates; unfortunately, working with functions defined via series greatly reduces their utility. The problem up to now has been the linkage between the Fourier transform of Φ and the Legendre series of ϕ.

In this paper, we provide such a linkage and then use it to obtain from Wendland's compactly supported RBFs [24, 25] a large class of explicit, locally supported SBFs on S^n whose Legendre series decay asymptotically at a prescribed algebraic rate. Moreover, these SBFs precisely meet the conditions required for the error estimates described above to hold. We do this in section 4.

In section 2 we discuss the relevant background material on SBFs and the various approximation results needed in the paper.

2. Background material.

2.1. The interpolation problem. Let $X := \{x_1, \ldots, x_N\}$ be a discrete set of distinct points on the n-sphere S^n. These will be the data sites for our interpolation problem. The *mesh norm* for X is

$$h_X = \sup_{y \in S^n} \inf_{x_j \in X} d(x_j, y),$$

where $d(x, y)$ is the geodesic (great circle) distance between the points x and y on S^n. The mesh norm measures the maximum distance any point on S^n can be from X. The *separation radius* is defined via

$$q_X = \frac{1}{2} \min_{j \neq k} d(x_j, x_k) \, .$$

This is half of the smallest geodesic distance between any two distinct points in X. It is easy to see that $h_X \geq q_X$; equality can hold only for a uniform distribution of points on S^1, the circle. The *mesh ratio*

$$\rho_X := h_X / q_X \geq 1$$

provides a measure of how uniformly points in X are distributed on S^n. In the case of the circle ($n = 1$), $\rho_X = 1$ means that the points are uniformly distributed. In all other cases, $\rho_X > 1$.

A continuous function $\phi : [-1, 1] \to R$ is said to be *positive definite* on S^n if the matrix $A_{j,k} := \phi(\cos(d(x_j, x_k)))$, x_j and x_k in X, is positive semidefinite for every possible finite set X of distinct points in S^n. It is said to be *strictly positive definite* on S^n if these matrices are all positive definite. Since $d(x, y)$ is the smaller of the angles between x and y on the great circle passing through these points, and since we may regard S^n as being embedded in R^{n+1} in the usual way, we have that $\cos(d(x, y)) = x \cdot y$, where $x \cdot y$ is the Euclidean inner product in R^{n+1} and ϕ induces a kernel $\phi(x \cdot y)$ that is in $C(S^n \times S^n)$. It follows that when ϕ is strictly positive definite, one can always solve this interpolation problem: Given the values $f(x_j)$, $j = 1, \ldots, N$, from sampling a continuous function on X, find coefficients $c_1, \ldots, c_N$ such that

$$I_X f(x) = \sum_k c_j \phi(x \cdot x_k)$$

agrees with f at the x_j's. This is so because the interpolation matrix $A_{j,k}$, being positive definite, is invertible.

Positive definite functions on spheres were introduced and characterized long ago by Schoenberg [22]. He showed that a function ϕ was positive definite if its expansion in Legendre polynomials in $n + 1$ variables,

$$\phi(x \cdot y) = \sum_{\ell=0}^{\infty} a_\ell P_\ell(n + 1, x \cdot y), \tag{2.1}$$

had all $a_\ell \geq 0$. We mention that Schoenberg actually used ultraspherical polynomials, which are proportional to the Legendre polynomials used here. The exact relation is [15, p. 33]

$$C_\ell^{\frac{n-1}{2}}(t) = \frac{\Gamma(\ell + n - 1)}{\Gamma(n - 1)\Gamma(\ell + 1)} P_\ell(n + 1; t) \, . \tag{2.2}$$

One can show that if $a_\ell > 0$ for all ℓ, then ϕ is strictly positive definite; see Xu and Cheney [26] and Ron and Sun [21]. When this holds for ϕ, we will call it an SBF. We remark that several recent review articles [1, 5, 12, 17] have dealt with such functions, and we refer the reader to them for further information.

2.2. Spherical harmonics, Sobolev spaces, and native spaces. The standard orthonormal basis of $L^2(S^n)$ is composed of spherical harmonics. Some basic facts will be described here, but we refer the reader to Müller's book [15] for further details. A spherical harmonic of order ℓ on S^n is the restriction to S^n of a homogeneous, harmonic polynomial in R^{n+1} of degree ℓ. We denote the space of spherical harmonics of order ℓ on S^n by V_ℓ and the dimension of V_ℓ by $N(n,\ell)$; this is given by [15, p. 4] as

$$N(n,0)=1 \quad \text{and} \quad N(n,\ell)=\frac{(2\ell+n-1)\Gamma(\ell+n-1)}{\Gamma(\ell+1)\Gamma(n)} \quad \text{for} \quad \ell\geq 1\,. \tag{2.3}$$

The space of spherical harmonics of order L or less will be denoted by $\mathcal{P}_L := \sum_{\ell=0}^{L} V_\ell$; it has dimension $\widetilde{N}(n,\ell)=N(n+1,\ell)$.

The space V_ℓ also has an intrinsic characterization; it is the eigenspace of the Laplace–Beltrami operator Δ on S^n corresponding to the eigenvalue ($\Delta Y_\ell+\lambda_\ell Y_\ell=0$)

$$\lambda_\ell=\ell(\ell+n-1), \quad \ell\geq 0.$$

Since Δ is a self-adjoint operator relative to the standard inner product,

$$\langle f,g\rangle=\int_{S^n} f(p)\overline{g(p)}d\sigma\,,$$

with $d\sigma$ being the volume element of S^n, the eigenspaces V_ℓ and $V_{\ell'}$, $\ell\neq\ell'$, are orthogonal relative to $\langle\cdot\,,\cdot\rangle$. As usual, one may choose an orthonormal basis for each V_ℓ, $\{Y_{\ell,m}\}_{m=1}^{N(n,\ell)}$. The collection of all the $Y_{\ell,m}$'s form an orthonormal basis for $L^2(S^n)$. Hence, for any function $f\in L^2(S^n)$, its associated (Fourier) series below converges in $L^2(S^n)$:

$$f=\sum_{\ell=0}^{\infty}\sum_{m=1}^{N(n,\ell)}\hat{f}_{\ell,m}Y_{\ell,m}\,, \quad \text{where } \hat{f}_{\ell,m}=\langle f,Y_{\ell,m}\rangle. \tag{2.4}$$

Such expansions (as in the case of periodic functions) can also be defined in a wider sense, namely for distributions on the sphere. Since S^n is compact, these distributions are the series (2.4) with tempered (i.e., polynomially bounded) coefficients. The Sobolev space $H_s(S^n)$ with real parameter s consists of all distributions f such that

$$\|f\|_{H_s}^2:=\sum_{\ell=0}^{\infty}\sum_{m=1}^{N(n,\ell)}(1+\lambda_\ell)^s|\hat{f}_{\ell,m}|^2=\|(I-\Delta)^{s/2}f\|_{L^2}^2<\infty. \tag{2.5}$$

See [10, section 1.7] and [6, Chapter II]. For later use, we define a related norm on functions in $C^{2k}(S^n)$; namely,

$$\|f\|_{2k}:=\max\{\|f\|,\|\Delta^k f\|\}, \quad f\in C^{2k}(S^n). \tag{2.6}$$

When $k=0$ this reduces to $\|f\|=\|f\|_{C(S^n)}$. In addition, we have

$$\|f\|_{H_{2k}}\leq 2^k\omega_n^{1/2}\|f\|_{2k}\,, \quad f\in C^{2k}(S^n), \tag{2.7}$$

where ω_n is the volume of S^n. This is easily established from $(1+\lambda_\ell)^{2k}\leq 1+(2^{2k}-1)\lambda_\ell^{2k}$ and that for all $g\in C(S^n)$ we have $\|g\|_{L^2}\leq\omega_n^{1/2}\|g\|$.

Let us now return to the expansion for ϕ given in (2.1). The famous addition theorem for spherical harmonics [15, Theorem 2] states that

$$P_\ell(n+1; x\cdot y) = \frac{\omega_n}{N(n,\ell)} \sum_{m=1}^{N(n,\ell)} Y_{\ell,m}(x)\overline{Y_{\ell,m}(y)}. \tag{2.8}$$

Using this in (2.1), we obtain the expansion

$$\phi(x\cdot y) = \sum_{\ell=0}^{\infty}\sum_{m=1}^{N(n,\ell)} \hat\phi(\ell) Y_{\ell,m}(x)\overline{Y_{\ell,m}(y)}, \quad \text{where } \hat\phi(\ell) := \frac{\omega_n}{N(n,\ell)} a_\ell\,. \tag{2.9}$$

Assume that ϕ is an SBF; i.e., $\hat\phi(\ell) > 0$ for all ℓ. One can now define the native space of ϕ to be

$$N_\phi := \left\{ f \in D'(S^n) \colon \ \|f\|_\Phi^2 = \sum_{\ell,k} |\hat f_{\ell,k}|^2/\hat\phi(\ell) < \infty \right\}. \tag{2.10}$$

As we noted in the introduction, if ϕ is smooth, then $\hat\phi(\ell)$ will decay rapidly and $1/\hat\phi(\ell)$ will grow. This means that functions in N_ϕ must have Fourier coefficients that decay rapidly for the sum in (2.10) to remain finite. This decay translates into smoothness for $f \in N_\phi$.

For such kernels, we have the following error estimates, which were established in [9] with improvements in [14] and [18].

Proposition 2.1. *Let X be any point set on S^n with mesh norm h_X, and let ϕ be an SBF, as in (2.9). If for some $\tau > \frac{n}{2}$ we have $\hat\phi(\ell) \le c(1+\lambda_\ell)^{-\tau}$ as $\ell \to \infty$, then for all $f \in N_\phi$ there is a constant C that is independent of X and f for which*

$$\|f - I_X f\| \le C h_X^{\tau - n/2} \|f\|_\phi.$$

Remark. The condition $\hat\phi(\ell) \le c(1+\lambda_\ell)^{-\tau}$ implies that $c/\hat\phi(\ell) \ge (1+\lambda_\ell)^\tau$, from which it immediately follows that $\|f\|_{H_\tau} \le c\|f\|_\phi$, so $N_\phi \subseteq H_\tau$. Conversely, if $\hat\phi(\ell) \ge c'(1+\lambda_\ell)^{-\tau'}$, with $\tau' \ge \tau$, then $\|f\|_{H_{\tau'}} \ge c'\|f\|_\phi$ and $N_\phi \supseteq H_{\tau'}$. In particular, if $\hat\phi(\ell) \sim (1+\lambda_\ell)^{-\tau}$, then $N_\phi = H_\tau$, and $\|\cdot\|_{H_\tau}$ and $\|\cdot\|_\phi$ are equivalent norms.

2.3. Approximation theorems. We will now collect several results concerning approximation of functions on S^n by spherical harmonics in $\mathcal{P}_L$, which are those of order L or less. These results were obtained by Pawelke [19, 20], who used two ideas from earlier works to obtain approximation results that we need here.

The first is the spherical mean of a function on S^n. The boundary of a spherical cap of radius $\arccos(h) < \pi$ and center x is the set $\{y \in S^n : x\cdot y = \cos(h)$, equivalently, $d(x,y) = \arccos(h)\}$, which is an $n-1$ dimensional sphere of radius $\sin h$. If $f \in C(S^n)$, then we define the spherical mean of f over $x\cdot y = \cos h$ to be

$$T_h f(x) := \frac{1}{\omega_{n-1}\sin^{n-1} h} \int_{x\cdot y = \cos h} f(y)\, d\sigma_x(y), \tag{2.11}$$

where $d\sigma_x$ is the volume element corresponding to $x\cdot y = \cos(\eta)$. Löfström and Peetre [11], in a study of approximation properties of orthogonal expansions, introduced similar operators.

The second is a spherical version of the modulus of continuity,

$$\omega(f,\varepsilon) := \sup_{0<h\leq\varepsilon} \|T_h f - f\|, \quad f \in C(S^n),\ \varepsilon > 0, \tag{2.12}$$

which is used below in estimating distance relative to $C(S^n)$ and in the proof in [20] for the Jackson inequality that follows it.

Theorem 2.2 (see [19, Satz 5.1] and [20, Satz 3.3]). *If $f \in C(S^n)$, then for $L = 1, 2, \ldots$ there is a constant M independent of both f and L for which*

$$\operatorname{dist}(f, \mathcal{P}_L) \leq M\omega(f; 1/L), \tag{2.13}$$

and for which

$$\operatorname{dist}(f, \mathcal{P}_L) \leq M^k L^{-2k} \|\Delta^k f\|, \quad k = 1, 2, \ldots,\ f \in C^{2k}(S^n). \tag{2.14}$$

In addition to the theorem above, we also need this Markov–Bernstein inequality.

Theorem 2.3 (see [20, Satz 3.6]). *If $P_L \in \mathcal{P}_L$, then*

$$\|\Delta P_L\| \leq D_n L^2 \|P_L\|\,,$$

where the constant D_n depends only on the dimension of the sphere S^n.

The remaining approximation results that we will make use of here have to do with the norm of iterates of Δ applied to best, and near-best, approximants from $\mathcal{P}_L$.

Proposition 2.4 (see [20, Satz 4.4]). *Let $f \in C^{2k}(S^n)$ and let P_L^* be a best approximant for f in $C(S^n)$; i.e., $\|f - P_L^*\| = \operatorname{dist}_\infty(f, \mathcal{P}_L)$. Then there exists a constant C independent of f and L for which*

$$\|\Delta^k P_L^*\| \leq C\|\Delta^k f\|.$$

Remark. Pawelke states his result in a weaker form. However, after observing that his sequence of operators $\{L_n\}$ commute with both Δ and the operators T_h and inspecting his proof, one obtains the result above.

There is an immediate, useful corollary to this proposition. In essence, it says that the theorem above applies to "near-best" approximants as well as the P_L^*.

Corollary 2.5. *Let $f \in C^{2k}(S^n)$ and let $P_L \in \mathcal{P}_L$, $L = 1, 2, \ldots,$ be a sequence of polynomials satisfying $\|f - P_L\| \leq K\operatorname{dist}(f, \mathcal{P}_L)$, with K independent of f and L. Then there is a constant R that is independent of f and L for which*

$$\|\Delta^k P_L\| \leq R\|\Delta^k f\|.$$

Proof. Observe that $\|\Delta^k P_L\| \leq \|\Delta^k(P_L - P_L^*)\| + \|\Delta^k P_L^*\|$. By iterating the Markov–Bernstein inequality in Theorem 2.3, we can bound the first term by the quantity $D_n^k L^{2k}\|P_L - P_L^*\|$. Moreover, since $\|P_L - P_L^*\| \leq \|P_L - f\| + \|f - P_L^*\|$, we have

$$\|\Delta^k(P_L - P_L^*)\| \leq D_n^k L^{2k}(1+K)\operatorname{dist}(f, \mathcal{P}_L)\,.$$

Applying the Jackson inequality (2.14) to the right side above, we arrive at the bound

$$\|\Delta^k(P_L - P_L^*)\| \leq D_n^k L^{2k}(1+K)M^k L^{-2k}\|\Delta^k f\| = (1+K)(MD_n)^k\|\Delta^k f\|.$$

By Proposition 2.4, we also have the inequality $\|\Delta^k P_L^*\| \leq C\|\Delta^k f\|$. It follows that

$$\|\Delta^k P_L\| \leq R\|\Delta^k f\|,$$

where $R = (1+K)(MD_n)^k + C$. □

3. Analysis of error in approximating by interpolants.

3.1. Error estimates. In this section, we derive error estimates for approximating a given function $f \in C^{2k}(S^n)$ by interpolants of the form

$$I_X f(x) := \sum_{x_j \in X} a_j \phi(x \cdot x_j)\,, \tag{3.1}$$

where ϕ is an SBF that satisfies $\hat{\phi}(\ell) \sim (1+\lambda_\ell)^{-\tau}$. As we noted at the end of section 2.2, this condition implies that $N_\phi = H_\tau$, with the norms being equivalent. Such estimates are already available in case $C^{2k}(S^n) \subset N_\phi$ or, what is equivalent here, $2k \geq \tau$; see [4, 7, 9]. Thus the primary cases of interest occur when $\tau > 2k$, so that $f \notin N_\phi$. Our main estimate addresses these cases. To prove it, we require constructing for every f in $C(S^n)$ spherical harmonics that both are near-best approximants to f from $\mathcal{P}_L$ and simultaneously interpolate f on the point set X. Precisely, we require this result.

THEOREM 3.1. *Let $X \subset S^n$ be a finite set of distinct points and let $\beta > 1$. If $L = \lceil \frac{(\beta+1)M}{(\beta-1)q_X} \rceil$, where M is as in Theorem 2.2, then for $f \in C(S^n)$ there exists a spherical harmonic $P_L \in \mathcal{P}_L$ that interpolates f on X and that satisfies*

$$\|f - P_L\| \leq (1+\beta)\,\mathrm{dist}(f, \mathcal{P}_L).$$

We will give the proof of Theorem 3.1 in section 3.2 below. We now state and prove our main estimate.

THEOREM 3.2. *Let ϕ be an SBF satisfying $\hat{\phi}(\ell) \sim (1+\lambda_\ell)^{-\tau}$ (equivalently, $N_\phi = H_\tau$), and suppose that $\tau \geq 2k > n/2$. If $f \in C^{2k}(S^n)$ and if $I_X f$ is given in* (3.1)*, then*

$$\|f - I_X f\| \leq C \rho_X^{\tau - 2k} h_X^{2k - n/2} \|f\|_{2k}\,,$$

where C is independent of f and X. Here, h_X and ρ_X are the mesh norm and mesh ratio for the set X, respectively.

Remark. If the point sets used are all quasi-uniformly distributed (i.e., $\rho_X \leq C$ for all X under consideration), then $\|f - I_X f\| \leq \tilde{C} h_X^{2k-n/2} \|f\|_{2k}$.

Proof. Note that since ϕ is an SBF, every spherical harmonic P is in the native space of ϕ. Thus for any spherical harmonic P,

$$\|f - I_X f\| \leq \|f - P\| + \|P - I_X P\| + \|I_X P - I_X f\|. \tag{3.2}$$

Moreover, if $P|_X = f|_X$, then $I_X f = I_X P$, and (3.2) becomes

$$\|f - I_X f\| \leq \|f - P\| + \|P - I_X P\|. \tag{3.3}$$

The main requirement for proving the estimate is having spherical harmonics that yield meaningful estimates in (3.3). This is in fact the content of Theorem 3.1. Indeed, choosing $\beta = 3$ there implies that we have a sequence of P_L with the following properties:

(A) $P_L \in \mathcal{P}_L$, where $L = \lceil 2M q_X^{-1} \rceil$, with M, which is independent of X, as in Theorem 2.2.
(B) $P_L|_X = f|_X$.
(C) $\|f - P_L\| \leq 4\,\mathrm{dist}(f, \mathcal{P}_L)$.

By (3.3), properties (A) through (C), and the Jackson inequality (2.14), we obtain

$$\begin{aligned}\|f - I_X f\| &\leq \|f - P_L\| + \|P_L - I_X P_L\| \\ &\leq 4 \operatorname{dist}(f, \mathcal{P}_L) + \|P_L - I_X P_L\| \\ &\leq 4M^k L^{-2k} \|\Delta^k f\| + \|P_L - I_X P_L\|. \end{aligned} \tag{3.4}$$

By the assumptions on ϕ, Proposition 2.1 holds and, since the norms $\|\cdot\|_\phi$ and $\|\cdot\|_{H_\tau}$ are equivalent, we can estimate the interpolation error for P_L on the right above via

$$\|P_L - I_X P_L\| \leq C h_X^{\tau - n/2} \|P_L\|_\phi \leq c h_X^{\tau - n/2} \|P_L\|_{H_\tau}. \tag{3.5}$$

Using the definition of the Sobolev norm given in (2.5) and the fact that P_L is a spherical harmonic of degree L, one can see that

$$\|P_L\|_{H_\tau} \leq (1 + \lambda_L)^{\tau/2 - k} \|P_L\|_{H_{2k}}\,.$$

In addition, employing the inequality in (2.7) to replace the norm $\|P_L\|_{H_{2k}}$, we have

$$\|P_L\|_{H_\tau} \leq 2^k \omega_n^{1/2} (1 + \lambda_L)^{\tau/2 - k} \|P_L\|_{2k}\,.$$

Next, from (C) we easily see that $\|P_L\| \leq 5\|f\|$, and from Corollary 2.5, we also have $\|\Delta^k P_L\| \leq R \|\Delta^k f\|$, so that $\|P_L\|_{2k} \leq \max\{5, R\} \|f\|_{2k}$ and, consequently,

$$\|P_L - I_X P_L\| \leq c h_X^{\tau - n/2} 2^k \omega_n^{1/2} (1 + \lambda_L)^{\tau/2 - k} \max\{5, R\} \|f\|_{2k}. \tag{3.6}$$

From (3.6), (3.4), and $\lambda_L = L(L + n - 1) \sim L^2$, we arrive at this bound on the interpolation error for f:

$$\|f - I_X f\| \leq \left(4M^k L^{-2k} + C_1 L^{\tau - 2k} h_X^{\tau - n/2}\right) \|f\|_{2k}\,.$$

Since $L \geq 1$, the last inequality can also be written as

$$\begin{aligned}\|f - I_X f\| &\leq \left(C_0 L^{n/2 - 2k} + C_1 h_X^{\tau - n/2} L^{\tau - 2k}\right) \|f\|_{2k} \\ &\leq \left(C_0 (h_X L)^{n/2 - 2k} + C_1 (h_X L)^{\tau - 2k}\right) h_X^{2k - n/2} \|f\|_{2k}\,.\end{aligned}$$

If we use $L = \lceil 2M/q_X \rceil = \lceil 2M\rho_X/h_X \rceil$ from (A), then we get

$$\|f - I_X f\| \leq \left(C_2 \rho_X^{n/2 - 2k} + C_3\, \rho_X^{\tau - 2k}\right) h_X^{2k - n/2} \|f\|_{2k}\,.$$

Finally, since $\rho_X \geq 1$ and $\tau > n/2$, it follows that

$$\|f - I_X f\| \leq C \rho_X^{\tau - 2k} h_X^{2k - n/2} \|f\|_{2k}\,. \qquad \square$$

A more general, but also more technical, result can be obtained by modifying the proof above.

Corollary 3.3. *Let $\tau' > \tau \geq 2k > n/2 + (\tau' - \tau)$. If $H_{\tau'} \subseteq N_\phi \subseteq H_\tau$, then, for f and $I_X f$ as in Theorem* 3.2*,*

$$\|f - I_X f\| \leq C \rho_X^{\tau' - 2k} h_X^{2k - n/2 - (\tau' - \tau)} \|f\|_{2k}$$

holds with C independent of f and X.

Proof (Sketch of proof). The condition that $N_\phi \subseteq H_\tau$ implies that the left inequality in (3.5) holds. On the other hand, $N_\phi \supseteq H_{\tau'}$ implies that the right inequality in (3.5) holds with $\|P_L\|_{H_{\tau'}}$ replacing $\|P_L\|_{H_\tau}$. Tracking the necessary changes through the rest of the proof then provides us with the desired estimate. $\square$

We close this section by pointing out that Corollary 3.3 shows that our estimate is optimized when $\tau' = \tau$ or, equivalently, when $N_\phi = H_\tau$. In section 4.2.2 we will obtain a family of compactly supported SBFs that satisfy this criterion.

3.2. Interpolants that are near-best approximants. We now will show that for a given $f \in C^{2k}(S^n)$ there exist spherical harmonics P_L satisfying properties (A) through (C) used in the proof above; i.e., we will prove Theorem 3.1. The key to producing such spherical harmonics is contained in the following proposition, which is an adaptation of a similar result [13, Theorem 2.1].

PROPOSITION 3.4. *Let $Z^* \subset C(S^n)^*$ be given by $Z^* = \operatorname{span}\{\delta_{x_j}\colon\ x_j \in X\}$, and let $\mathcal{V}$ be a finite dimensional subspace of $C(S^n)$. If for every $z^* \in Z^*$ and some $\beta > 1$, β independent of z^*,*

$$\|z^*\|_{C(S^n)^*} \le \beta \|z^*|_{\mathcal{V}}\|_{\mathcal{V}^*}\,,$$

then for $f \in C(S^n)$ there exists $v_f \in \mathcal{V}$ for which $f|_X = v_f|_X$ and $\|f - v\| \le (1+\beta)\operatorname{dist}(f,\mathcal{V})$.

Proof. Let $v_\star$ be a best approximant to f from $\mathcal{V}$, so that

$$\|f - v_\star\| = \operatorname{dist}(f,\mathcal{V}),$$

and set $e := f - v_\star$. Let the restriction map $S\colon\ Z^* \to Z^*|_{\mathcal{V}}$ be given by $S(z^*) = z^*|_{\mathcal{V}}$ for every $z^* \in Z^*$. Since $\|z^*\| \le \beta\|z^*|_{\mathcal{V}}\|$, S is both one-to-one and onto the image space $S(Z^*) \subset \mathcal{V}^*$. Moreover $\|S^{-1}\| \le \beta$, where $S^{-1}\colon\ S(Z^*) \to Z^*$. Viewing e as an element of Z^{**} (i.e., as a functional on Z^*), we have

$$\langle e, z^*\rangle = \langle S^*(S^*)^{-1}e, z^*\rangle = \langle (S^*)^{-1}e, Sz^*\rangle\,,$$

where we used the fact that $S = S^{**}$. Note that $(S^*)^{-1}e \in (S(Z^*))^*$, where $S(Z^*) \subset \mathcal{V}^*$. By the Hahn–Banach theorem, $(S^*)^{-1}e$ extends in a norm-preserving manner to $v_e \in \mathcal{V}^{**} = \mathcal{V}$. Thus $\langle e, z^*\rangle = \langle f - v_\star, z^*\rangle = \langle z^*, v_e\rangle$ for all $z^* \in Z^*$ and

$$\begin{aligned}\|v_e\| &= \|(S^*)^{-1}e\| \le \|S^{-1}\|\|e\| \le \beta\|e\| \\ &= \beta\|f - v^*\|.\end{aligned}$$

Setting $v_f := v_\star + v_e$ gives an element in $\mathcal{V}$ for which $f|_X = v_f|_X$ and

$$\|f - v_f\| \le \|f - v^*\| + \|v_e\| \le (1+\beta)\ \operatorname{dist}(f,\mathcal{V}),$$

which completes the proof. □

The point of this proposition is that we have reduced the problem of finding interpolants that are near-best approximants to one of finding the ratio of norms of linear functionals. To estimate these ratios, we will construct a norm-attaining function for $z^* := \sum_{x_j\in X} c_j\delta_{x_j} \in Z^*$ and then approximate that function with spherical harmonics. We will show that the function ζ, which is given as

$$\zeta(x) := \sum_{x_j\in X} \operatorname{sgn}(c_j)\left(1 - \frac{d(x,x_j)}{q_X}\right)_+\,, \tag{3.7}$$

has the required properties.

LEMMA 3.5. *Let the function $\zeta(x)$ be defined by* (3.7). *Then, ζ is continuous on S^n and satisfies these properties:*

(i) $\|\zeta\| = 1$.

(ii) $z^*(\zeta) = \|z^*\|$.

(iii) $\omega(\zeta,\varepsilon) \le \frac{\varepsilon}{q_X}, 0 < \varepsilon \le \frac{\pi}{2}$, *where $\omega(\zeta, \varepsilon)$ is given in* (2.12).

Proof. Continuity is obvious. Also, $\zeta(x) = 0$ unless $d(x, x_j) < q_X$ for some $x_j \in X$. Moreover, on $d(x, x_j) \leq q_X$, we have

$$\zeta(x) = \operatorname{sgn}(c_j)\left(1 - \frac{d(x, x_j)}{q_X}\right),$$

so $|\zeta(x)| = 1 - \frac{d(x,x_j)}{q_X} \leq |\zeta(x_j)| = 1$. Since x_j is arbitrary, we see that $|\zeta(x)| \leq 1$ for all $x \in S^n$. Since $\zeta(x_j) = 1$ for all $x_j \in X$, $\|\zeta\| = 1$.

To establish (iii), we first need to calculate certain directional derivatives. We will work on $d(x_j, x) < q_X$, with x_j regarded as the north pole of S^n. In that case, we have $d(x, x_j) = \theta$, the colatitude of X, so

$$\zeta(x) = \operatorname{sgn}(c_j)\left(1 - \frac{\theta}{q}\right), \qquad d(x, x_j) = \theta < q_X.$$

For any geodesic $x(s)$, with $x(0) = x$ and s the arclength, we have

$$\frac{d}{ds}\zeta(x(s)) = -\operatorname{sgn}(c_j) \cdot \frac{1}{q}\frac{d\theta}{ds}.$$

In standard spherical coordinates,

$$1 = \left(\frac{d\theta}{ds}\right)^2 + \text{positive terms},$$

so $\left|\frac{d\theta}{ds}\right| \leq 1$. It follows that

$$\left|\frac{d\zeta}{ds}(x(0))\right| \leq \frac{1}{q_X}.$$

If $d(x, x_j) > q_X$ for all $x_j \in X$, then $\zeta(x) = 0$ in a neighborhood of x, and $\frac{d\zeta}{ds}(x(0)) = 0$. The only difficulty occurs where $d(x, x_j) = q$. Although ζ is not continuously differentiable at such points, the directional derivatives exist and again we have $|\frac{d\zeta}{ds}| \leq \frac{1}{q_X}$. (If we pass through such a point, $\frac{d\zeta}{ds}$ will have a jump discontinuity.) Our main consequence is that if *any* points x and y on S^n are joined by a geodesic $x(s)$, with $x(0) = x$ and $x(\varepsilon) = y$, then

$$|\zeta(x) - \zeta(y)| = \left|\int_0^\varepsilon \frac{d\zeta}{ds}(x(s))ds\right| \leq \int_0^\varepsilon \frac{ds}{q_X} = \frac{\varepsilon}{q_X}.$$

From (2.12), the spherical modulus of continuity $\omega(\zeta, \varepsilon)$ is

$$\omega(\zeta, \varepsilon) = \sup_{0<\eta\leq\epsilon} \|T_\eta\zeta - \zeta\|,$$

where $T_\eta\zeta$ is the spherical mean defined in (2.11). It is easy to show that $T_\eta c = c$ for any constant c, so that

$$T_\eta\zeta(x) - \zeta(x) = \frac{1}{\omega_{n-1}\sin^{n-1}(\eta)}\int_{x\cdot y=\cos(\eta)} \big(\zeta(y) - \zeta(x)\big)d\sigma(y).$$

Consequently

$$|T_\eta\zeta(x)-\zeta(x)| \le \frac{1}{\omega_{n-1}\sin^{n-1}(\eta)}\int_{x\cdot y=\cos\eta}\big|\zeta(x)-\zeta(y)\big|d\sigma(y)$$
$$\le \frac{\eta}{q_X}\cdot\frac{\int d\sigma(y)}{\omega_{n-1}\sin^{n-1}(\eta)} = \frac{\eta}{q_X}.$$

From this, one sees that $\omega(\zeta,\varepsilon)\le\frac{\varepsilon}{q_X}$, as we claimed. □

With minor modifications, a similar result can be proven for a Riemannian manifold. One need work only in the usual normal coordinates obtained from the exponential map. These can be used to define all the quantities involved.

Knowing the properties of ζ, we are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. Apply Theorem 2.2 and Lemma 3.5 to ζ to obtain the existence of a polynomial $P_\zeta\in\mathcal{P}_L$ such that $\|P_\zeta-\zeta\|\le M/(q_XL)$. If, in addition, we assume that $L=\lceil\frac{(\beta+1)M}{(\beta-1)q_X}\rceil$, we see that

$$\|P_\zeta-\zeta\|\le\frac{\beta-1}{\beta+1}.$$

Furthermore, since $\|\zeta\|=1$, we also have that

$$\|P_\zeta\|\le\frac{2\beta}{\beta+1}.$$

Suppose that $z^*\in\operatorname{span}\{\delta_{x_1},\dots,\delta_{x_N}\}$ and that $\|z^*\|=1$. From Lemma 3.5 and the identity $z^*(\zeta-P_\zeta)+z^*(P_\zeta)=1$, we see that

$$z^*(P_\zeta)\ge 1-|z^*(\zeta-P_\zeta)|\ge 1-\frac{\beta-1}{\beta+1}=\frac{2}{\beta+1}.$$

Consequently,

$$\|z^*\|=1\le\frac{\beta+1}{2}z^*(P_\zeta)\le\frac{\beta+1}{2}\|z^*|_{\mathcal{P}_L}\|\cdot\|P_\zeta\|\le\frac{\beta+1}{2}\cdot\frac{2\beta}{\beta+1}\|z^*|_{\mathcal{P}_L}\|=\beta\|z^*|_{\mathcal{P}_L}\|.$$

The theorem is then an immediate consequence of Proposition 3.4, with $\mathcal{V}=\mathcal{P}_L$. □

Remark. In the case of the circle T, one can work with Sobolev spaces of degree k instead of $2k$ and use derivatives in place of Δ. Also the verification that $\|P_L\|_{2k}\le C\|f\|_{2k}$ follows easily by applying the appropriate theorems from [2, Chap. 7, sect. 2].

4. Restrictions of RBFs in R^{n+1} to S^n.

In our main estimate, Theorem 3.2, we assumed that $\hat\phi(\ell)\sim(1+\lambda_\ell)^{-\tau}$. Earlier, in a remark at the end of section 2.2, we pointed out that this assumption amounted to $N_\phi=H_\tau$, with the norms on the spaces being equivalent. In this section, we will exhibit SBFs that satisfy this property. Indeed, we will show that the compactly supported RBFs constructed by Wendland [24, 25] restrict (as kernels) to locally supported SBFs with native spaces that coincide with Sobolev spaces.

4.1. Legendre coefficients.

We suppose that Φ is a positive definite radial function defined on R^{n+1} and having the Fourier representation

$$\Phi(x)=\frac{1}{(2\pi)^{n+1}}\int_{R^{n+1}}\hat\Phi(|\xi|)e^{i\xi\cdot x}d\xi,\quad \hat\Phi\ge 0,\ \hat\Phi(|\cdot|)\in L^1(R^{n+1}). \tag{4.1}$$

Of course, if Φ has such a representation, then Φ is rotationally invariant and is thus a function of $|x|$ only. The corresponding convolution kernel $\Phi(x-y)$ is, when $|x|=1$ and $|y|=1$, a function of $|x-y| = \sqrt{2(1-x\cdot y)}$. Consequently, the restriction $\Phi(x-y)|_{x,y\in S^n}$ is a function of $x\cdot y$. We may therefore define the function

$$\phi(x\cdot y) := \Phi(x-y)|_{x,y\in S^n}\,. \tag{4.2}$$

Note that $\phi(x\cdot y)$ inherits being positive definite from Φ, and so it has the expansion given in (2.9). Our immediate goal is to use (2.9) to express $\hat\phi(\ell)$ in terms of $\hat\Phi$. In addition to being an interesting formula in its own right, it will be the key to relating native spaces for Φ and ϕ. We have the following result.

THEOREM 4.1. *Let ϕ defined in* (4.2) *have the expansion* (2.9). *Then $\hat\phi(\ell)$ from* (2.9) *is given by*

$$\hat\phi(\ell) = \int_0^\infty t\hat\Phi(t)J^2_{\ell+\frac{n-1}{2}}(t)dt, \tag{4.3}$$

where $J_\nu(t)$ is the usual Bessel function of the first kind and of order ν.

Proof. From (2.9), we have that

$$\hat\phi(\ell) = \int_{S^n}\int_{S^n}\phi(x\cdot y)\overline{Y_{\ell,m}(x)}Y_{\ell,m}(y)d\sigma(x)d\sigma(y)\,.$$

Substituting (4.1) for Φ in (4.2), inserting the resultant expression for ϕ into the previous equation, and interchanging integrals, we obtain

(4.4)
$$\hat\phi(\ell) = \frac{1}{(2\pi)^{n+1}}\int_{R^{n+1}}\hat\Phi(|\xi|)\bigg(\int_{S^n}e^{-i\xi\cdot y}Y_{\ell,m}(y)d\sigma(y)\int_{S^n}e^{i\xi\cdot x}\,\overline{Y_{\ell,m}(x)}d\sigma(x)\bigg)d\xi\,.$$

From Watson's book [23, sect. 11.5, eq. (2)], we have that

$$e^{it\cos\theta} = 2^\nu\Gamma(\nu)\sum_{\ell=0}^\infty(\nu+\ell)i^\ell\frac{J_{\nu+\ell}(t)}{t^\nu}C_\ell^\nu(\cos\theta),$$

where C_ℓ^ν is a Gegenbauer polynomial. We choose $\nu=\frac{n-1}{2}$ and replace the Gegenbauer polynomials by the corresponding Legendre polynomials using (2.2), where we also use the expression for $N(n,\ell)$ in (2.3):

$$\begin{aligned} 2^{\frac{n-1}{2}}\Gamma(\tfrac{n-1}{2})(\ell+\tfrac{n-1}{2})C_\ell^{\frac{n-1}{2}}(\cos\theta) &= 2^{\frac{n+1}{2}-1}\Gamma(\tfrac{n+1}{2})N(n,\ell)P_\ell(n+1;\cos\theta)\\ &= (2\pi)^{\frac{n+1}{2}}\tfrac{N(n,\ell)}{\omega_n}P_\ell(n+1;\cos\theta)\,. \end{aligned}$$

Employing this in the expansion for $e^{it\cos\theta}$, we arrive at

$$e^{it\cos\theta} = \frac{(2\pi)^{\frac{n+1}{2}}}{t^{\frac{n-1}{2}}}\sum_{\ell=0}i^\ell J_{\ell+\frac{n-1}{2}}(t)\frac{N(n,\ell)}{\omega_n}P_\ell(n+1;\cos\theta)\,.$$

Now choose $t=|\xi|$, $\eta=\xi/t$, $\cos\theta=\eta\cdot x$, and use the addition theorem (2.8); the result is

$$e^{it\eta\cdot x} = \frac{(2\pi)^{\frac{n+1}{2}}}{t^{\frac{n-1}{2}}}\sum_{\ell=0}i^\ell J_{\ell+\frac{n-1}{2}}(t)\sum_{m=1}^{N(n,\ell)}Y_{\ell,m}(x)\overline{Y_{\ell,m}(\eta)}\,. \tag{4.5}$$

From this, we can compute the inner integrals in (4.4); in particular, for the integral in x, we have

$$\int_{S^n} e^{it\eta\cdot x}\overline{Y_{\ell,m}(x)}d\sigma(x) = \frac{(2\pi)^{\frac{n+1}{2}}i^\ell}{t^{\frac{n-1}{2}}}J_{\ell+\frac{n-1}{2}}(t)\overline{Y_{\ell,m}(\eta)}\,. \tag{4.6}$$

The integral in y is simply the complex conjugate of that in x. Hence, it follows that

$$\hat{\phi}(\ell) = \frac{1}{(2\pi)^{n+1}}\int_0^\infty \hat{\Phi}(t)\int_{S^n}\left|\frac{(2\pi)^{\frac{n+1}{2}}i^\ell}{t^{\frac{n-1}{2}}}J_{\ell+\frac{n-1}{2}}(t)\overline{Y_{\ell,m}(\eta)}\right|^2 d\sigma(\eta)t^n dt\,.$$

Doing the integral over S^n and simplifying the integrand results in (4.3), which completes the proof. □

Having an explicit relationship between the Fourier transform of an RBF and the Legendre coefficients provides an important relationship between native spaces.

COROLLARY 4.2. *Let Φ and Ψ be positive definite radial functions on R^{n+1}, and suppose that $\hat{\Psi}$ and $\hat{\Phi}$ are strictly positive and satisfy $\hat{\Phi} \le c\hat{\Psi}$; then, for all $\ell \ge 0$, we have $0 < \hat{\phi}(\ell) \le c\hat{\psi}(\ell)$ and $N_\phi \subseteq N_\psi$.*

Proof. Apply the previous theorem. □

Earlier, we defined an SBF to be a positive definite function on S^n with the additional property that $\hat{\phi}(\ell) > 0$ for all $\ell \ge 0$. This condition implies that the standard interpolation matrices are positive definite and hence invertible. Ron and Sun [21] showed that this condition was sufficient but not necessary for the positive definiteness of the interpolation matrices. The condition is, however, necessary and sufficient for doing generalized Hermite interpolation [4, 16]. Up to now, it was only known that if Φ was an RBF on R^{n+1}, then, given very mild conditions (see [3]), $\Phi(x-y)|_{S^{n-1}}$ was an SBF; that is, one had to drop down two dimensions. One knew only that $\Phi(x-y)|_{S^n}$ was positive definite, but *not* that it was an SBF.

COROLLARY 4.3. *Let Φ be a nontrivial positive definite radial function having the form given in* (4.1). *Then, the restriction $\phi(x\cdot y) := \Phi(x-y)|_{x,y\in S^n}$ is an SBF.*

Proof. We first note that $\hat{\Phi}$ is positive on a set of nonzero measure, for Φ would vanish identically otherwise. Let $\mathcal{M}$ be this set. Suppose that ϕ is not an SBF, so that for some ℓ we have $\hat{\phi}(\ell) = 0$. (Of course, ϕ is positive definite, so that $\hat{\phi}(\ell) \ge 0$.) From (4.3), we see that on $\mathcal{M}$

$$\int_{\mathcal{M}} t\hat{\Phi}(t)J^2_{\ell+\frac{n-1}{2}}(t)dt = 0.$$

Since the measure of $\mathcal{M}$ is positive, and since $\hat{\Phi}(t) > 0$ on $\mathcal{M}$, it follows that the continuous function $J_{\ell+\frac{n-1}{2}}(t) \equiv 0$ on $\mathcal{M}$, and, consequently, the entire function $t^{-\ell+\frac{n-1}{2}}J_{\ell+\frac{n-1}{2}}(t)$ vanishes identically on $\mathcal{M}$. Since $\mathcal{M}$ is uncountable, the entire function vanishes identically, as does the Bessel function $J_{\ell+\frac{n-1}{2}}$, which is false. Hence, $\hat{\phi}(\ell) > 0$ for all ℓ, and ϕ is an SBF. □

4.2. SBFs with Sobolev spaces for native spaces. A positive definite radial function Φ on R^{n+1} has its native space N_Φ equivalent to a Sobolev space $H_s(R^{n+1})$ if its R^{n+1}-Fourier transform $\hat{\Phi}(|\xi|)$ satisfies the bounds

$$c(1+t^2)^{-s} \le \hat{\Phi}(t) \le C(1+t^2)^{-s}, \quad 0 \le t \in R\,. \tag{4.7}$$

We will assume that $\hat{\Phi}$ satisfies these bounds, and, in order to keep $\hat{\Phi}$ in $L^1(R^{n+1})$, we will also require that s be strictly larger than $(n+1)/2$.

We want to determine whether $\phi(x\cdot y) = \Phi(x-y)|_{x,y\in S^n}$ belongs to some Sobolev space on S^n, given that $\hat{\Phi}$ satisfies the bounds in (4.7). Define Ψ_s via

$$\Psi_s(x) := \frac{1}{(2\pi)^{n+1}}\int_{R^{n+1}} \frac{e^{ix\cdot\xi}}{(1+|\xi|^2)^s}d\xi\,, \quad x\in R^{n+1},$$

and let $\psi_s(x\cdot y) = \Psi_s(x-y)|_{x,y\in S^n}$. By Corollary 4.2 and the bounds in (4.7), the native spaces for ϕ and ψ_s are the same, with the norms being equivalent. The question of what, if any, Sobolev space ψ_s belongs to can be answered by using (4.3) to obtain the large ℓ behavior of $\hat{\psi}_s(\ell)$, which is explicitly given by

$$\hat{\psi}_s(\ell) = \int_0^\infty \frac{tJ_\nu^2(t)}{(1+t^2)^s}dt, \text{ where } \nu := \ell + \frac{n-1}{2} \text{ and } s > \frac{n+1}{2}. \tag{4.8}$$

Specifically, we have this proposition.

Proposition 4.4. *If for some $\tau(s,n) > 0$ we have $\hat{\psi}_s(\ell) \sim \ell^{-2\tau}$ as $\ell\to\infty$, then $N_{\psi_s} = H_\tau(S^n)$. Moreover, if Φ is such that (4.7) holds, then we also have $N_\phi = H_\tau(S^n)$.*

Proof. The asymptotic behavior of $\hat{\psi}_s(\ell)$ implies that for all $\ell \geq 0$,

$$(1+\lambda_\ell)^\tau \hat{\psi}_s(\ell) \sim 1, \quad \ell\to\infty,$$

since $\lambda_\ell = \ell(\ell+n-1) = \ell^2(1+\mathcal{O}(\ell^{-1}))$. Now, by Corollary 4.3, ψ_s is an SBF and so $\hat{\psi}_s(\ell) > 0$ for all $\ell \geq 0$. Consequently, the asymptotic statement in the equation above holds for all $\ell\geq 0$, and thus the native space of ψ_s coincides with $H_\tau(S^n)$. The statements concerning N_ϕ then follow from the bound (4.7) and Corollary 4.2. □

4.2.1. Large $\boldsymbol{\ell}$ asymptotics of $\boldsymbol{\hat{\psi}_s(\ell)}$. We now turn to finding the large ℓ asymptotic behavior of $\hat{\psi}_s(\ell)$, at least for some s. We begin by evaluating the integral (4.8) in terms of hypergeometric functions [23, section 4.4],

$${}_pF_q(a_1,a_2,\ldots,a_p\,;b_1,b_2,\ldots,b_q\,;z) := \sum_{r=0}^\infty \frac{(a_1)_r(a_2)_r\cdots(a_p)_r z^r}{(b_1)_r(b_2)_r\cdots(b_q)_r r!}\,, \tag{4.9}$$

where Pochhammer's symbol $(\lambda)_r := \lambda(\lambda+1)\cdots(\lambda+r-1)$ when $r\geq 1$ and $(\lambda)_0 := 1$.

Lemma 4.5. *If s and ν are as in (4.8) and $\nu - s$ is not an integer, then*

$$\begin{aligned}\hat{\psi}_s(\ell) = {}& \frac{2\Gamma(\nu+1-s)\Gamma(s-\frac{1}{2})}{\pi^{\frac{1}{2}}\Gamma(s)\Gamma(\nu+s)} \times \Bigg({}_1F_2(s-\tfrac{1}{2}; s+\nu, s-\nu; 1) \\ & + \frac{\pi^{\frac{3}{2}}(\nu+1-s)\Gamma(\nu+s)\csc\big(\pi(\nu-s)\big)}{2^{2\nu}\Gamma^2(\nu+2-s)\Gamma(\nu+1)\Gamma(s-\frac{1}{2})}\, {}_1F_2(\nu+\tfrac{1}{2}; \nu+2-s, 2\nu+1; 1)\Bigg).\end{aligned} \tag{4.10}$$

Proof. We follow Watson [23, section 13.61], who sketched two methods for doing integrals involving products of Bessel functions, specifically including integrals of the type in (4.8). The most direct for us is to first use [23, eq. (1), sect. 5.43] to express $J_\nu^2(t)$ as

$$J_\nu^2(t) = \frac{2}{\pi}\int_0^{\pi/2} J_{2\nu}(2t\cos\theta)d\theta,$$

then insert it in (4.8), and finally use Fubini's theorem to interchange integrals. This results in

$$\hat{\psi}_s(\ell) = \frac{2}{\pi}\int_0^{\pi/2}\int_0^\infty \frac{tJ_{2\nu}(2t\cos\theta)}{(1+t^2)^s}dtd\theta\,.$$

The integral over t was done in [23, sect. 13.6, eq. (1)]; with our parameters, it has the form

$$\int_0^\infty \frac{tJ_{2\nu}(2t\cos\theta)}{(1+t^2)^s}dt = \frac{\Gamma(\nu+1)\Gamma(s-1-\nu)}{2\Gamma(s)\Gamma(2\nu+1)}\cos^{2\nu}(\theta)\,{}_1F_2(\nu+1;\nu+2-s,2\nu+1;\cos^2(\theta)) + \frac{\Gamma(\nu+1-s)}{2\Gamma(s+\nu)}\cos^{2s-2}(\theta)\,{}_1F_2(s;s+\nu,s-\nu;\cos^2(\theta))\,.$$

We are now left with evaluating two integrals of the form

$$\int_0^{\pi/2}\cos^\mu(\theta)\,{}_1F_2(a_1;b_1,b_2;\cos^2(\theta))d\theta\,.$$

If we let $u=\cos^2(\theta)$, then such integrals transform to

$$2\int_0^1 u^{\frac{\mu-1}{2}}(1-u)^{-\frac{1}{2}}\,{}_1F_2(a_1;b_1,b_2;u)du\,.$$

This integral may be found in Gradshteyn and Ryzhik [8, sect. 7.512, eq. 12]:

$$\int_0^1 u^{\frac{\mu-1}{2}}(1-u)^{-\frac{1}{2}}\,{}_1F_2(a_1;b_1,b_2;u)du = \frac{\pi^{1/2}\Gamma(\frac{\mu+1}{2})}{\Gamma(1+\frac{\mu}{2})}\,{}_2F_3(\tfrac{\mu+1}{2},a_1;1+\tfrac{\mu}{2},b_1,b_2;1)\,.$$

Using this result, we have that

$$\begin{aligned}\hat{\psi}_s(\ell) =& A\,{}_2F_3(\nu+\tfrac{1}{2},\nu+1;\nu+1,\nu+2-s,2\nu+1;1)\\ &+B\,{}_2F_3(s-\tfrac{1}{2},s;s,s+\nu,s-\nu;1),\end{aligned}$$

where A and B are the accumulated factors and are given by

$$A=\frac{2\Gamma(s-1-\nu)\Gamma(\nu+\frac{1}{2})}{\pi^{\frac{1}{2}}\Gamma(s)\Gamma(2\nu+1)} \quad\text{and}\quad B=\frac{2\Gamma(\nu+1-s)\Gamma(s-\frac{1}{2})}{\pi^{\frac{1}{2}}\Gamma(s)\Gamma(\nu+s)}\,.$$

Using the "cancellation property" for the hypergeometric functions, namely,

$${}_{p+1}F_{q+1}(c,a_1,\dots,a_p\,;c,b_1,\dots,b_p;z) = {}_pF_q(a_1,\dots,a_p\,;b_1,\dots,b_p;z),$$

we arrive at

$$\hat{\psi}_s(\ell) = A\,{}_1F_2(\nu+\tfrac{1}{2};\nu+2-s,2\nu+1;1)+B\,{}_1F_2(s-\tfrac{1}{2};s+\nu,s-\nu;1)\,.$$

Using $\Gamma(1-z)\Gamma(z)=\pi\csc(\pi z)$ and the duplication formula, $2^{2z-1}\Gamma(z)\Gamma(z+\frac{1}{2}) = \sqrt{\pi}\Gamma(2z)$, we can rewrite A as

$$A=\frac{2^{1-2\nu}\pi\csc\big(\pi(\nu-s)\big)}{\Gamma(s)\Gamma(\nu+2-s)\Gamma(\nu+1)}\,.$$

Inserting this into the previous equation for $\hat{\psi}_s(\ell)$ and factoring out B, we obtain the expression in (4.10). □

The restriction that $\nu - s$ cannot be an integer is really unnecessary; the expression in (4.10) has a removable singularity for such values, and $\hat{\psi}_s(\ell)$ can be found for them by taking limits.

We now turn to the large ℓ asymptotics of $\hat{\psi}_s(\ell)$. Doing this for fixed, *arbitrary* $s > \frac{n+1}{2}$ appears to be quite difficult, owing to the removable singularities that occur in (4.10) when $\nu - s$ is an integer. These disappear in one very important case, namely when $\nu - s$ is an odd multiple of $\frac{1}{2}$. This case is important because the compactly supported positive definite radial functions introduced by Wendland [24, 25] satisfy it (see section 4.2.2). For such functions we have $s = \frac{n+1}{2} + j + \frac{1}{2}$, where j is a positive integer, and so

$$\nu - s = \ell - j - \tfrac{3}{2} \quad \text{and} \quad \nu + s = \ell + n + j + \tfrac{1}{2}\,. \tag{4.11}$$

With this choice of s, we have the following asymptotic formula.

PROPOSITION 4.6. *As $\ell \to \infty$,*

$$\hat{\psi}_{\frac{n+1}{2}+j+\frac{1}{2}}(\ell) = \frac{2\Gamma(\frac{n+1}{2}+j)}{\sqrt{\pi}\Gamma(\frac{n+1}{2}+j+\frac{1}{2})}\ell^{-2j-n-1}\big(1+\mathcal{O}(\ell^{-1})\big)\,.$$

Proof. With $s = \frac{n+1}{2} + j + \frac{1}{2}$ and the values of $\nu \pm s$ from (4.11), equation (4.10) for $\hat{\psi}_s(\ell)$ becomes

$$\begin{aligned}\hat{\psi}_{\frac{n+1}{2}+j+\frac{1}{2}}(\ell) = {} & \frac{2\Gamma(\frac{n+1}{2}+j)\Gamma(\ell-j-\frac{1}{2})}{\sqrt{\pi}\Gamma(\frac{n+1}{2}+j+\frac{1}{2})\Gamma(\ell+n+j+\frac{1}{2})} \\ & \times\Bigg({}_1F_2(\tfrac{n+1}{2}+j;\ell+n+j+\tfrac{1}{2},j+\tfrac{3}{2}-\ell;1) \\ & \pm\frac{\pi^{\frac{3}{2}}(\ell-j-\frac{1}{2})\Gamma(\ell+n+j+\frac{1}{2})}{2^{2\ell+n-1}\Gamma^2(\ell-j+\frac{1}{2})\Gamma(\ell+\frac{n+1}{2})\Gamma(\frac{n+1}{2}+j)}\,{}_1F_2(\ell+\tfrac{n}{2};\ell-j+\tfrac{1}{2},2\ell+n;1)\Bigg)\,,\end{aligned}$$

where the "±" comes from $\csc(\pi(\nu - s))$. We will take care of terms in reverse order. By the definition of the hypergeometric function in (4.9), it is easy to see that

$${}_1F_2(\ell+\tfrac{n}{2};\ell-j+\tfrac{1}{2},2\ell+n;1) \le \sum_{r=0}^{\infty}\frac{1}{r!} \le e.$$

The term multiplying this hypergeometric is roughly $\mathcal{O}((\ell!)^{-2})$. Thus overall the term decays faster than any power of ℓ. Again from (4.9),

$$\begin{aligned}\big|\,{}_1F_2(\tfrac{n+1}{2}+j;\ell+n+j+\tfrac{1}{2},j+\tfrac{3}{2}-\ell;1)-1\big| &\le \textstyle\sum_{r=1}^{\infty}\frac{(\frac{n+1}{2}+j)_r}{(\ell+n+j+\frac{1}{2})_r|(j+\frac{3}{2}-\ell)_r|r!} \\ &\le 4\textstyle\sum_{r=1}^{\infty}\frac{(\frac{n+1}{2}+j)_r}{(\ell+n+j+\frac{1}{2})_r r!} \\ &\le 4\frac{\frac{n+1}{2}+j}{\ell+n+j+\frac{1}{2}}\textstyle\sum_{r=1}^{\infty}\frac{(\frac{n+1}{2}+j+1)_{r-1}}{(\ell+n+j+\frac{1}{2})_{r-1}r!} \\ &\le 4(e-1)\frac{\frac{n+1}{2}+j}{\ell+n+j+\frac{1}{2}}\,.\end{aligned}$$

Consequently, the first term in parentheses is $1 + \mathcal{O}(\ell^{-1})$. Since the second term decays faster than any power of ℓ, we have that the two terms taken together behave like $1+\mathcal{O}(\ell^{-1})$. Using standard properties of the Gamma function, one has that the term multiplying the parentheses is

$$\frac{2\Gamma(\frac{n+1}{2}+j)}{\sqrt{\pi}\Gamma(\frac{n+1}{2}+j+\frac{1}{2})}\times\ell^{-2j-n-1}(1+\mathcal{O}(\ell^{-1}))\,.$$

The proposition follows on observing that $(1+\mathcal{O}(\ell^{-1}))\times(1+\mathcal{O}(\ell^{-1}))=1+\mathcal{O}(\ell^{-1})$. $\square$

We conjecture that for any $s>\frac{n+1}{2}$ we will have $\hat{\psi}_s(\ell)\sim\ell^{-2s+1}$ as $\ell\to\infty$.

4.2.2. Locally supported SBFs with $N_\phi = H_s$. As we mentioned earlier, in [24] Wendland introduced a class of compactly supported RBFs, and in [25] he explored their properties, showing in particular that their Fourier transforms satisfy the bounds in (4.7). On R^d, these functions have the form

$$\Phi_{d,j}(x) := \begin{cases} p_{d,j}(\|x\|_2) & \text{if } \|x\|_2\le 1,\\ 0 & \text{if } \|x\|_2 > 1,\end{cases}$$

where $x\in R^d$ and $p_{d,j}$ is a polynomial of degree $\lfloor\frac{d}{2}\rfloor+3j+1$; in addition, $\Phi_{d,j}$ is in $C^{2j}(R^d)$ [25, Cor. 2.3] and satisfies (4.7) with $s=\frac{d}{2}+j+\frac{1}{2}$ [25, Thm. 2.1], where $j\ge 1$ and $d=1,2,\ldots$. Simply choosing $d=n+1$ puts s in the form $s=\frac{n+1}{2}+j+\frac{1}{2}$. By Proposition 4.6 and Proposition 4.4, it follows that, with $\tau=\frac{n+1}{2}+j$, the native space for $\phi_{n,j}$ and the Sobolev space H_τ are equivalent. The specific result is this.

THEOREM 4.7. *Let $\phi_{n,j}(x\cdot y) := \Phi_{\frac{n+1}{2},j}(x-y)|_{x,y\in S^n}$. Then $\phi_{n,j}$ is a C^{2j} spherical basis function for which $N_{\phi_{n,j}} = H_{\frac{n+1}{2}+j}$.*

We point out that the support of $\phi_{n,d}$ can be adjusted by scaling $\Phi_{\frac{n+1}{2},j}$. This will not change any of $\phi_{n,d}$'s essential properties, and so the theorem above holds for it as well.

One interesting feature is that, in restricting to S^n, the native space of $\Phi_{\frac{n+1}{2},j}$ changes from $H_s(R^{n+1})$ to $H_{s-\frac{1}{2}}(S^n)$, with $s=\frac{n+1}{2}+j+\frac{1}{2}$. This loss of "$\frac{1}{2}$ a derivative" is familiar from the theory of Sobolev spaces and traces of functions. The trace operator restricts a function in a Sobolev space to a codimension 1 surface, such as an embedded sphere in Euclidean space. The trace of f belongs to a Sobolev space of order $\frac{1}{2}$ smaller than the one f is in [10, Chapter 1, section 8]. We conjecture that the result we have obtained is a special case of a more general one that applies to manifolds; namely, we conjecture that if the native space of Φ coincides with $H_s(\Omega)$, for an $n+1$ dimensional manifold Ω, then the restriction ϕ to an n dimensional surface Σ will have $H_{s-\frac{1}{2}}(\Sigma)$ for a native space.

REFERENCES

[1] W. Cheney, *Approximation using positive definite functions*, in Approximation Theory VIII, Vol. 1: Approximation and Interpolation, C. K. Chui and L. L. Schumaker, eds., World Scientific Publishing, Singapore, 1995, pp. 145–168.

[2] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, Grundlehren Math. Wiss. 303, Springer-Verlag, Berlin, 1993.

[3] N. Dyn, F. J. Narcowich, and J. D. Ward, *A framework for interpolation and approximation on Riemannian manifolds*, in Approximation Theory and Optimization, M. D. Buhmann and A. Iserles, eds., Cambridge University Press, Cambridge, UK, 1997, pp. 133–144.

[4] N. Dyn, F. J. Narcowich, and J. D. Ward, *Variational principles and Sobolev-type estimates for generalized interpolation on a Riemannian manifold*, Constr. Approx., 15 (1999), pp. 175–208.

[5] G. E. Fasshauer and L. L. Schumaker, *Scattered data fitting on the sphere*, in Mathematical Methods for Curves and Surfaces II, M. Dæhlen, T. Lyche, and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 1998, pp. 117–166.

[6] P. B. Gilkey, *The Index Theorem and the Heat Equation*, Publish or Perish, Boston, MA, 1974.

[7] M. von Golitschek and W. A. Light, *Interpolation by polynomials and radial basis functions*, Constr. Approx., 17 (2001), pp. 1–18.

[8] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, San Diego, CA, 1980.

[9] K. Jetter, J. Stöckler, and J. D. Ward, *Error estimates for scattered data interpolation*, Math. Comp., 68 (1999), pp. 743–747.

[10] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications,* Vol. I, Springer-Verlag, New York, 1972.

[11] J. Löfström and J. Peetre, *Approximation theorems connected with generalized translation*, Math. Ann., 181 (1969), pp. 255–268.

[12] H. N. Mhaskar, F. J. Narcowich, and J. D. Ward, *Representing and analyzing scattered data on spheres*, in Multivariate Approximation and Applications, N. Dyn, D. Leviaton, D. Levin, and A. Pinkus, eds., Cambridge University Press, Cambridge, UK, 2001, pp. 44–72.

[13] H. N. Mhaskar, F. J. Narcowich, N. Sivakumar, and J. D. Ward, *Approximation with interpolatory constraints*, Proc. Amer. Math. Soc., 130 (2002), pp. 1355–1364.

[14] T. M. Morton and M. Neamtu, *Error bounds for solving pseudodifferential equations on spheres by collocation with zonal kernels*, J. Approx. Theory, 114 (2002), pp. 242–268.

[15] C. Müller, *Spherical Harmonics*, Lecture Notes in Math. 17, Springer-Verlag, Berlin, 1966.

[16] F. J. Narcowich, *Generalized Hermite interpolation and positive definite kernels on a Riemannian manifold*, J. Math. Anal. Appl., 190 (1995), pp. 165–193.

[17] F. J. Narcowich, *Recent developments in approximation via positive definite functions*, in Approximation IX, Vol. II: Computational Aspects, C. K. Chui and L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 1998, pp. 221–242.

[18] F. J. Narcowich, R. Schaback, and J. D. Ward, *Approximation in Sobolev spaces by kernel expansions*, J. Approx. Theory, 114 (2002), pp. 70–83.

[19] S. Pawelke, *Ein Satz vom Jacksonschen Typ für algebraischen Polynome*, Acta Sci. Math. (Szeged), 33 (1972), pp. 323–336.

[20] S. Pawelke, *Über die Approximationsordnung bei Kugelfunktionen und algebraischen Polynomen*, Tôhoku Math. J. (2), 24 (1972), pp. 473–486.

[21] A. Ron and X. Sun, *Strictly positive definite functions on spheres on spheres in Euclidean spaces*, Math. Comp., 65 (1996), pp. 1513–1530.

[22] I. J. Schoenberg, *Positive definite functions on spheres*, Duke Math. J., 9 (1942), pp. 96–108.

[23] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, London, 1966.

[24] H. Wendland, *Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree*, Adv. Comput. Math., 4 (1995), pp. 389–396.

[25] H. Wendland, *Error estimates for interpolation by compactly supported radial basis functions of minimal degree*, J. Approx. Theory, 93 (1998), pp. 258–272.

[26] Y. Xu and E. W. Cheney, *Strictly positive definite functions on spheres*, Proc. Amer. Math. Soc., 116 (1992), pp. 977–981.

[27] J. Yoon, *L_p-error estimates for ‘shifted’ surface spline interpolation on Sobolev space*, Math. Comp., to appear.

SIAM J. MATH. ANAL.
Vol. 33, No. 6, pp. 1411–1429

LARGE TORSIONAL OSCILLATIONS IN A SUSPENSION BRIDGE: MULTIPLE PERIODIC SOLUTIONS TO A NONLINEAR WAVE EQUATION[*]

K. S. MOORE[†]

Abstract. We consider a forced nonlinear wave equation on a bounded domain which, under certain physical assumptions, models the torsional oscillation of the main span of a suspension bridge. We use Leray–Schauder degree theory to prove that, under small periodic external forcing, the undamped equation has multiple periodic solutions. To establish this multiplicity theorem, we prove an abstract degree theoretic result that can be used to prove multiplicity of solutions for more general operators and nonlinearities.

Using physical constants from the engineers' reports of the collapse of the Tacoma Narrows Bridge, we solve the damped equation numerically and observe that multiple periodic solutions exist and that whether the span oscillates with small or large amplitude depends only on its initial displacement and velocity. Moreover, we observe that the qualitative properties of our computed solutions are consistent with the behavior observed at Tacoma Narrows on the day of its collapse.

Key words. nonlinear wave equation, torsional oscillations, suspension bridge

AMS subject classification. 35B10

PII. S0036141001388099

1. Introduction.

For over sixty years, scientists in many disciplines have struggled to explain the dramatic and finally destructive torsional oscillations of the Tacoma Narrows Bridge that preceded its collapse in 1940. The recent article in [16], which describes the forty year effort to control the behavior of the Deer Isle Bridge in Maine, and the closing in June, 2000, of the Millennium Bridge in London [18] testify to the fact that the problem of controlling suspension bridge oscillations remains unsolved.

We argue that the nonlinearity inherent in the equations of motion drives the unpredictable behavior observed on the Tacoma Narrows and other suspension bridges. Theoretical and numerical evidence to support this claim for the vertical, torsional, and traveling wave motion of suspension bridges can be found in [3], [4], [5], [6], [7] and [9], [10], [11], [12], [13], [14], [15].

In [9] and [10], the authors proposed an ODE model for the torsional motion of a horizontal cross section of the main span of a suspension bridge and proved the existence of multiple periodic solutions. Using physical constants from the engineers' reports of the Tacoma Narrows collapse, the authors investigated this model numerically and demonstrated that under small external forcing, the cross section may ultimately settle down to small or large amplitude periodic torsional oscillation, depending only on the initial torsional displacement and velocity of the cross section.

In this paper, we extend this analysis to the entire length of the main span of the bridge. More specifically, in section 2 we propose a PDE model (the forced sine-Gordon equation on a bounded domain) for the torsional motion along the length of the center span. In section 3, we prove that, under certain physical assumptions, the equation has multiple periodic weak solutions. Similar results exist for the vertical motion of the center span [6], [13].

[*] Received by the editors April 19, 2001; accepted for publication (in revised form) December 28, 2001; published electronically July 9, 2002.

http://www.siam.org/journals/sima/33-6/38809.html

[†] Department of Mathematics, University of Michigan, Ann Arbor, MI 48109-1109 (ksmoore@umich.edu).

We then investigate these solutions numerically. In section 4, we examine the bifurcation properties of periodic solutions to the equation via numerical continuation algorithms. We find that, under small external forcing, the damped equation has three periodic solutions, one of small amplitude and two of large amplitude. Moreover, we see that bifurcation from single to multiple solutions occurs for small forcing.

In section 5, we use finite difference methods to approximate periodic solutions. As in [9], we demonstrate that under small external forcing, the center span may oscillate periodically with small or large amplitude, depending only on its initial displacement and velocity. Moreover, we observe that the qualitative properties such as amplitude, frequency, and nodal structure of our computed solutions are consistent with the behavior observed at Tacoma Narrows on the day of its collapse.

2. The model. We treat the center span of the bridge as a beam of length L and width $2l$ suspended by cables (see Figure 2.1). Consider the horizontal cross section of mass m located at position x along the length of the span. We treat this cross section as a rod of length $2l$ and mass m suspended by cables. Let $y(x,t)$ denote the *downward* distance of the center of gravity of the rod *from the unloaded state* and let $\theta(x,t)$ denote the angle of the rod from horizontal at time t (see Figure 2.1).

We assume that the cables do not resist compression, but resist elongation according to Hooke's Law with spring constant K; i.e., the force exerted by the cable is proportional to the elongation in the cable with proportionality constant K. In

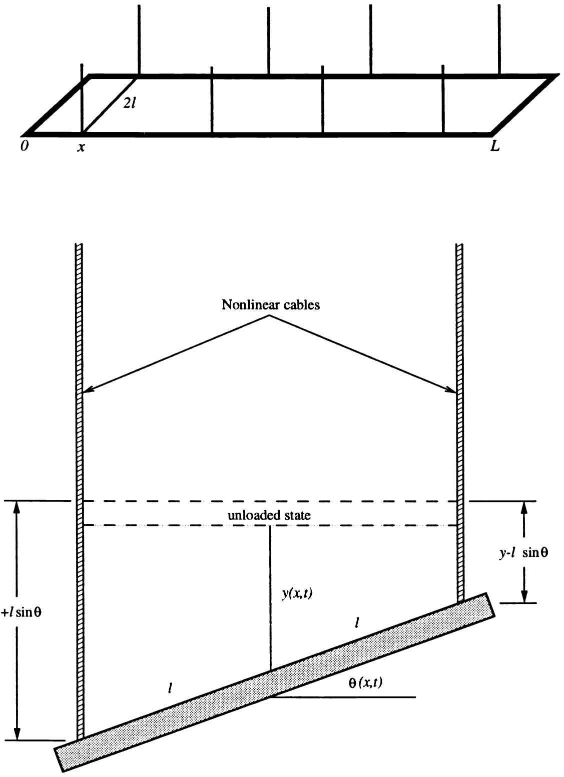


FIG. 2.1. *A simple model of the center span and its horizontal cross section.*

Figure 2.1 we see that the extension in the right-hand cable is $(y - l\sin\theta)$, and hence the force exerted by the right-hand cable is

$$\left\{\begin{array}{ll} -K(y - l\sin\theta), & y - l\sin\theta \geq 0 \\ 0, & y - l\sin\theta < 0 \end{array}\right. = -K(y - l\sin\theta)^+,$$

where $u^+ = \max\{u, 0\}$. Similarly, the extension in the left-hand cable is $(y + l\sin\theta)$, and the force exerted by the left-hand cable is $-K(y + l\sin\theta)^+$. Then the torsional and vertical motion of the span satisfy

$$\text{(2.1)} \quad \left\{\begin{array}{c} \theta_{tt} - \varepsilon_1\theta_{xx} = \frac{3K}{ml}\cos\theta[(y - l\sin\theta)^+ - (y + l\sin\theta)^+] - \delta\theta_t + h_1(x,t) \\ y_{tt} + \varepsilon_2 y_{xxxx} = -\frac{K}{m}[(y - l\sin\theta)^+ + (y + l\sin\theta)^+] - \delta y_t + g + h_2(x,t) \\ \theta(0,t) = \theta(L,t) = y(0,t) = y(L,t) = y_{xx}(0,t) = y_{xx}(L,t) = 0 \end{array}\right\},$$

where $\varepsilon_1, \varepsilon_2$ are physical constants related to the flexibility of the beam, δ is the damping constant, h_1 and h_2 are external forcing terms, and g is the acceleration due to gravity. The spatial derivatives describe the restoring force that the beam exerts, and the time derivatives θ_t and y_t represent the force due to friction. The boundary conditions reflect the fact that the ends of the span are hinged.

We study coupled systems of this form in [11], [12], and [15]. However, throughout this paper we assume that the cables never lose tension; i.e., we assume that $(y \pm l\sin\theta) \geq 0$ and hence $(y \pm l\sin\theta)^+ = (y \pm l\sin\theta)$. In this case, we see that the equations (2.1) become uncoupled, and the torsional and vertical motions satisfy

$$\text{(2.2)} \quad \left\{\begin{array}{c} \theta_{tt} - \varepsilon_1\theta_{xx} = -\frac{6K}{m}\cos\theta\sin\theta - \delta\theta_t + h_1(x,t) \\ \theta(0,t) = \theta(L,t) = 0 \end{array}\right\}$$

and

$$\text{(2.3)} \quad \left\{\begin{array}{c} y_{tt} + \varepsilon_2 y_{xxxx} = -\frac{2K}{m}y - \delta y_t + g + h_2(x,t) \\ y(0,t) = y(L,t) = y_{xx}(0,t) = y_{xx}(L,t) = 0 \end{array}\right\},$$

respectively.

We observe that (2.2) is the damped, forced, sine-Gordon equation, which arises in many applications. We study equations of this form throughout the paper.

3. A multiplicity theorem. In this section, we consider the questions of existence and multiplicity of continuous, periodic, weak solutions u in a subspace of $\mathcal{L}^2$ to equations of the form (2.2).

Let $\Omega = (0,\pi) \times (0,\pi)$ and define

$$\mathcal{H} = \{u \in \mathcal{L}^2(\Omega) | u(x,t) = u(\pi - x, t), u(x,t) = u(x, \pi - t), \\ u \text{ is } \pi \text{ periodic in } t\}.$$

For $u \in \mathcal{H}$, let $\|u\| = \|u\|_{\mathcal{L}^2} = (\int_\Omega |u|^2 dA)^{\frac{1}{2}}$. Define $L_T u = u_{tt} - u_{xx}$. Using $\cos u \sin u = \frac{1}{2}\sin 2u$, changing variables, removing the damping term, and imposing boundary and periodicity conditions, we rewrite (2.2) as

$$\text{(3.1)} \quad \left\{\begin{array}{c} L_T u + b\sin u = \varepsilon h(x,t) \\ u(0,t) = u(\pi,t) = 0 \\ u(x,0) = u(x,\pi), u_t(x,0) = u_t(x,\pi) \end{array}\right\}.$$

Observe that the eigenvalues and corresponding eigenfunctions of L_T with the appropriate boundary conditions are

$$\left\{ \begin{array}{l} \lambda_{mn} = (2n+1)^2 - 4m^2 \\ \phi_{mn} = \cos(2mt)\sin((2n+1)x) \end{array} \right\}, \tag{3.2}$$

where $m, n = 0, 1, 2, \ldots$. Because we are restricted to the subspace $\mathcal{H}$ of $\mathcal{L}^2$, L_T^{-1} exists, is compact, and $\|L_T^{-1}\| = 1$.

Definition 3.1. *We say that $u \in \mathcal{H}$ is a solution to* (3.1) *if*

$$u = L_T^{-1}(\varepsilon h - b\sin u). \tag{3.3}$$

Theorem 3.2. *Let $h \in \mathcal{H}$ with $\|h\| \leq 1$, and let $b \in (3, 7)$. Then there exists $\varepsilon_0 > 0$ such that if $|\varepsilon| < \varepsilon_0$,* (3.1) *has at least two solutions in $\mathcal{H}$.*

We use Leray–Schauder degree theory to prove Theorem 3.2 in section 3.2; however, to establish this result, we first establish a general degree theoretic result in section 3.1. Finally, in section 3.3, we prove that the solutions to (3.1) are continuous.

3.1. Preservation of Leray–Schauder degree under Gâteaux differentiation. To establish the existence of multiple periodic solutions to (3.1), we use Leray–Schauder degree theory to prove the existence of multiple zeros of a related operator T_1. To compute the degree of T_1, we continuously deform it to a linear operator T_0, the Gâteaux derivative of T_1, and compute its degree via a direct calculation.

It is not difficult to show that, under the appropriate hypotheses, the homotopy property of Leray–Schauder degree ensures that the degree of an operator T_1 is preserved as T_1 is continuously deformed to its Fréchet derivative. However, the nonlinear term in (3.1), $f(u) = \sin u$, is not Fréchet differentiable in $\mathcal{L}^2$ at $u = 0$.

Motivated by the result and arguments in [13], in Theorem 3.3 we show that, under certain conditions on the nonlinear term f and the differential operator L, Leray–Schauder degree is indeed preserved under homotopy from the operator T_1 to its Gâteaux derivative T_0. This result can be used to establish multiplicity of solutions to equations of the form (3.1) for more general nonlinearities $f(u)$ and differential operators L.

Theorem 3.3. *Let I_1, I_2 be open, bounded intervals in $\mathbf{R}$, and define $Q := I_1 \times I_2$. Let $\mathcal{B}$ be a subspace of $\mathcal{L}^p(Q), p \geq 1$, and define $\|u\| := \|u\|_{\mathcal{L}^p}$. Consider the problem*

$$Lu + f(u) = \varepsilon h(x, t), \tag{3.4}$$

where L, f, and h satisfy the following:

(H1) *L^{-1} is compact;*

(H2) *$\|L^{-1}\| \leq 1$;*

(H3) *$f(0) = 0$;*

(H4) *f is Lipschitz with Lipschitz constant M;*

(H5) *$h \in \mathcal{B}$ and $\|h\| \leq 1$;*

(H6) *the Gâteaux derivative $df(0, u)$ exists and satisfies $df(0, u) = \rho u$, where $\rho > 0$ and $-\rho$ is not an eigenvalue of L.*

Define $T_0 : \mathcal{B} \to \mathcal{B}$ by $T_0(u) = u + \rho L^{-1}(u)$ and $T_1 : \mathcal{B} \to \mathcal{B}$ by $T_1(u) = u - L^{-1}(\varepsilon h - f(u))$. Then for ε sufficiently small, there exists $\gamma > 0$ such that $deg(T_1, B_\gamma(0), 0) = deg(T_0, B_\gamma(0), 0)$.

Proof. For $\lambda \in [0, 1]$, define $T_\lambda : \mathcal{B} \to \mathcal{B}$ by

$$T_\lambda(u) = u + (1 - \lambda)\rho L^{-1}(u) + \lambda L^{-1}(f(u) - \varepsilon h(x, t)). \tag{3.5}$$

The homotopy property of degree ensures that $deg(T_\lambda, B_r(0), 0)$ is constant provided that $0 \notin T_\lambda(\partial B_r(0))$ for $\lambda \in [0,1]$. We will show then that for all $\lambda \in [0,1]$ there exists $\gamma > 0$ such that the solution u to $T_\lambda(u) = 0$ satisfies $\|u\| \neq \gamma$.

Observe that $u = 0$ is the only zero of T_0 since, by (H6), $-\rho$ is not an eigenvalue of L. Fix $\lambda \in (0,1]$ and suppose that $u \neq 0$ solves $T_\lambda(u) = 0$. Set $\|u\| = \tilde{\gamma}_\lambda > 0$. We will show that $\tilde{\gamma}_\lambda$ is bounded below by some $\gamma_\lambda > 0$.

Note that u solves

$$Lu + \rho u = \lambda(\rho u - f(u) + \varepsilon h(x,t)) \tag{3.6}$$

and hence

$$Lu = (\lambda - 1)\rho u + \lambda(\varepsilon h(x,t) - f(u)), \tag{3.7}$$

and invoking (H4) and (H5) we have

$$\begin{aligned} \|Lu\| &\leq \rho\|u\| + \varepsilon + \|f(u)\| \\ &\leq \rho\|u\| + \varepsilon + M\|u\| \\ &= (\rho + M)\tilde{\gamma}_\lambda + \varepsilon. \end{aligned}$$

Therefore, $u \in \overline{L^{-1}B_{(\rho+M)\tilde{\gamma}_\lambda+\varepsilon}(0)}$, which is compact by (H1). Set $\psi = \frac{u}{\tilde{\gamma}_\lambda}$. Then $\|\psi\| = 1$ and there exists a compact set K with $\psi \in K$.

Since u solves (3.6), we have

$$\|u + \rho L^{-1}u\| = \lambda\|L^{-1}(\rho u - f(u) + \varepsilon h)\|. \tag{3.8}$$

Denote the left- and right-hand sides of (3.8) by LHS and RHS, respectively. Since $-\rho$ is not an eigenvalue of L, we have $L\psi + \rho\psi \neq 0$ and hence

$$\inf_{\psi \in K} \|\psi + L^{-1}\rho\psi\| = \alpha > 0, \tag{3.9}$$

and therefore for our u we have

$$LHS = \|u + L^{-1}\rho u\| \geq \alpha\tilde{\gamma}_\lambda. \tag{3.10}$$

Now considering RHS, by (H2) and (H5), we have

$$\begin{aligned} RHS &\leq \lambda\|\rho u - f(u) + \varepsilon h\| \\ &\leq \lambda[\varepsilon + \|\rho u - f(u)\|]. \end{aligned}$$

We claim that if ε and $\tilde{\gamma}_\lambda$ are sufficiently small, $RHS < \alpha\tilde{\gamma}_\lambda$, which contradicts (3.10). To establish this, we must first prove the following lemma.

LEMMA 3.4. *Let $f, \mathcal{B}, \rho$ be as in the statement of Theorem 3.3, and let $K \subset \mathcal{B}$ be compact. Then there exists a function $\delta : (0,\infty) \to (0,\infty)$ such that*

(L1) $\|\rho\eta\psi - f(\eta\psi)\| \leq \eta\delta(\eta)$,

(L2) $\delta(\eta) \to 0$ *as* $\eta \to 0$

hold for all $\psi \in K$ and $\eta > 0$.

Proof. Define $\delta : (0,\infty) \to (0,\infty)$ by

$$\delta(\eta) = \max_{\psi \in K} \left\|\rho\psi - \frac{1}{\eta}f(\eta\psi)\right\|, \tag{3.11}$$

and note that (L1) above is satisfied.

To show that (L2) holds, we must show that $\|\rho\psi - \frac{1}{\eta}f(\eta\psi)\| \to 0$ uniformly on K as $\eta \to 0$. Define $f_\eta : K \to \mathbf{R}$ by

$$f_\eta(\psi) = \left\| \rho\psi - \frac{1}{\eta} f(\eta\psi) \right\|. \tag{3.12}$$

To show that $f_\eta \to 0$ uniformly on K, we will show that $f_\eta(\psi) \to 0$ for each $\psi \in K$ and that the family $\mathcal{F} := \{f_\eta\}$ is equicontinuous on K. Choose $\psi \in K$. If $\psi = 0$, then $f_\eta(\psi) = 0$, so assume $\psi \neq 0$. By (H6), we have $\frac{d}{dt}f(t\psi) \mid_{t=0} = \rho\psi$ and hence, given $\tilde{\varepsilon} > 0$, for η sufficiently small, using (H3) and (H6), we have

$$\left\| \frac{1}{\eta} f(\eta\psi) - \rho\psi \right\| < \tilde{\varepsilon}. \tag{3.13}$$

To see that the family $\mathcal{F} = \{f_\eta\}$ is equicontinuous on K, choose $\tilde{\varepsilon} > 0$ and $\psi, \tilde{\psi} \in K$. Using (H4), we have

$$\begin{aligned} |f_\eta(\psi) - f_\eta(\tilde{\psi})| &= \left| \left\| \rho\psi - \frac{1}{\eta} f(\eta\psi) \right\| - \left\| \rho\tilde{\psi} - \frac{1}{\eta} f(\eta\tilde{\psi}) \right\| \right| \\ &\leq \left\| \rho(\psi - \tilde{\psi}) - \frac{1}{\eta}(f(\eta\psi) - f(\eta\tilde{\psi})) \right\| \\ &\leq \rho\|\psi - \tilde{\psi}\| + \frac{1}{\eta} M \|\eta\psi - \eta\tilde{\psi}\| \\ &= (\rho + M)\|\psi - \tilde{\psi}\| < \tilde{\varepsilon}, \end{aligned}$$

provided $\|\psi - \tilde{\psi}\| < \delta := \frac{\tilde{\varepsilon}}{\rho + M}$.

Since $\{f_\eta\}$ are equicontinuous on K and converge pointwise on K, we have that f_η converge uniformly on K, and hence (L2) holds. □

Returning now to the proof of the theorem and invoking the above lemma, we have that

$$\begin{aligned} RHS &\leq \lambda[\varepsilon + \|\rho u - f(u)\|] \\ &\leq \lambda[\varepsilon + \|\rho\tilde{\gamma}_\lambda\psi - f(\tilde{\gamma}_\lambda\psi)\|] \\ &\leq \lambda[\varepsilon + \tilde{\gamma}_\lambda\delta(\tilde{\gamma}_\lambda)]. \end{aligned}$$

Assume now that $\varepsilon < \frac{1}{2}\alpha\tilde{\gamma}_\lambda$. Take $\lambda = 1$. Since $\delta \to 0$, there exists γ such that $x < \gamma$ implies that $\delta(x) < \frac{1}{2}\alpha$. Moreover, for any $\lambda \in (0,1)$, if $x < \gamma$, $\lambda\delta(x) < \frac{1}{2}\alpha$. If $\tilde{\gamma}_\lambda = \|u\| < \gamma$, we have $RHS < \alpha\tilde{\gamma}_\lambda$. But this contradicts (3.10). Thus we conclude that $0 \notin T_\lambda(\partial B_\gamma(0))$ and therefore $deg(T_1, B_\gamma(0), 0) = deg(T_0, B_\gamma(0), 0)$. □

3.2. The proof of Theorem 3.2. Note that by (3.2) and by our choice of $b \in (3, 7)$, $-b$ is *not* an eigenvalue of L_T; moreover, there are no negative eigenvalues of L_T between $\lambda_{10} = -3$ and $\lambda_{21} = -7$.

Define $T_1 : \mathcal{H} \to \mathcal{H}$ by

$$T_1(u) = u - L_T^{-1}(\varepsilon h - b\sin(u))$$

and note that zeros of T_1 correspond to solutions of (3.1). To prove the theorem, we will show

(D1) there exists $R_0 > 0$ such that for $R > R_0$, $deg(T_1, B_R(0), 0) = 1$ and

(D2) there exists $\gamma \in (0, R_0)$ such that $deg(T_1, B_\gamma(0), 0) = -1$.

Then, since $deg(T_1, B_\gamma(0), 0) \neq 0$, there exists a zero of T_1 (i.e., a solution of (3.1)) in $B_\gamma(0)$. Moreover, by the additivity property of degree, $deg(T_1, B_R(0)\backslash\overline{B_\gamma(0)}, 0) \neq 0$ and hence (3.1) has a second solution in the annulus $B_R(0)\backslash\overline{B_\gamma(0)}$.

To establish (D1), define

$$T_\beta u = u - \beta L_T^{-1}(\varepsilon h - b\sin(u))$$

for $\beta \in [0, 1]$, and note that this definition of T_1 is consistent with our previous definition. Note also that T_0 is simply the identity map; hence, for any $R > 0$ we have $deg(T_0, B_R(0), 0) = 1$. The homotopy property of degree ensures that $deg(T_\beta, B_R(0), 0)$ is constant provided that $0 \notin T_\beta(\partial B_R(0))$ for all $\beta \in [0, 1]$.

Fix $\beta \in [0, 1]$ and suppose $u \in \mathcal{H}$ solves $T_\beta u = 0$. We will show that u is bounded above by some $R_0 > 0$ and that this bound is independent of β.

Since $T_\beta u = 0$, we have

$$\begin{aligned}\|u\| &= \beta\|L_T^{-1}(\varepsilon h - b\sin u)\| \leq \beta[\varepsilon_0 + b\|\sin u\|] \\ &\leq [\varepsilon_0 + bm(\Omega)^{\frac{1}{2}}] = [\varepsilon_0 + b\sqrt{2}\pi] < R_0\end{aligned}$$

if we choose $R_0 > \varepsilon_0 + b\sqrt{2}\pi$.

Thus, for $R > R_0$ we have

$$deg(T_1, B_R(0), 0) = deg(T_0, B_R(0), 0) = 1, \tag{3.14}$$

and (D1) above holds.

To establish (D2), let $\varepsilon < \varepsilon_0$; we will determine the value of ε_0 later. For $\mu \in [0, 1]$ define

$$T_\mu u = u + (1 - \mu)L_T^{-1}(bu) - \mu L_T^{-1}(\varepsilon h - b\sin u),$$

and note again that this definition of T_1 is consistent with our previous definitions. We will again apply the homotopy property of degree (via Theorem 3.3) and a standard degree calculation to show that for some $\gamma > 0$

$$deg(T_1, B_\gamma(0), 0) = deg(T_0, B_\gamma(0), 0) = -1.$$

Observe that for $L = L_T$ and $f(u) := b\sin u$, hypotheses (H1)–(H5) of Theorem 3.3 are satisfied. To verify hypothesis (H6), we need to show that

$$df(0, u) = bu. \tag{3.15}$$

By definition of the Gâteaux derivative,

$$\begin{aligned}df(0, u) &= \frac{d}{dt}f(0 + tu)\mid_{t=0} \\ &= \lim_{h\to 0}\frac{f((t+h)u) - f(tu)}{h}\Big|_{t=0} \\ &= \lim_{h\to 0}\frac{b\sin(hu)}{h}.\end{aligned}$$

We will show that the limit above (in $\mathcal{H}$) is bu.

Note first that in $\mathbf{R}$ we have

$$\lim_{h\to 0}\frac{\sin(hu)}{h} = \lim_{h\to 0}\frac{\sin(hu)}{h}\frac{u}{u} = u$$

and hence

$$\left|\frac{\sin(hu)}{h} - u\right|^2 \to 0$$

as $h \to 0$. Invoking the convexity of w^2, we have

$$\left|\frac{\sin(hu)}{h} - u\right|^2 \leq 4\left[\frac{1}{2}\left|\frac{\sin(hu)}{h}\right|^2 + \frac{1}{2}|u|^2\right] \leq 4u^2.$$

Since $u \in \mathcal{L}^2$, $|\frac{\sin(hu)}{h} - u|^2$ is dominated in $\mathcal{L}^1$; thus by the dominated convergence theorem,

$$\left\|\frac{b\sin(hu)}{h} - bu\right\| \to 0$$

as $h \to 0$; therefore (3.15) holds. Moreover, by (3.2) and our choice of b, $-b$ is not an eigenvalue of L_T; therefore hypothesis (H6) of Theorem 3.3 holds. Thus, by Theorem 3.3, for sufficiently small $\gamma, \varepsilon > 0$, we have

$$deg(T_1, B_\gamma(0), 0) = deg(T_0, B_\gamma(0), 0). \tag{3.16}$$

Finally, we will show that

$$deg(T_0, B_\gamma(0), 0) = deg(I + bL_T^{-1}, B_\gamma(0), 0) = -1.$$

Consider the finite dimensional subspace $\mathcal{M}_N$:=span$\{\phi_{mn}\}_1^N$ of $\mathcal{H}$ and recall that, by compactness, bL_T^{-1} can be approximated in operator norm by the operators $B_N : \mathcal{M}_N \to \mathcal{M}_N$ given by

$$B_N(u) = b\sum_{m,n=1}^{N}\frac{c_{mn}}{\lambda_{mn}}\phi_{mn}.$$

By definition of Leray–Schauder degree, for N sufficiently large,

$$\begin{aligned} deg(T_0, B_\gamma(0), 0) &= deg(I + B_N, B_\gamma(0) \cap \mathcal{M}_N, 0) \\ &= \textstyle\sum_{u\in(I+B_N)^{-1}(0)} signJ_{I+B_N}(u), \end{aligned} \tag{3.17}$$

where $J_\phi(u)$ is the Jacobian determinant of ϕ at u.

Since $I + B_N$ can be identified with an $N^2 \times N^2$ diagonal matrix whose entries are $1 + \frac{b}{\lambda_{mn}}$, we have

$$deg(I + B_N, B_\gamma(0) \cap \mathcal{M}_N, 0) = sign\prod_{m,n=1}^{N}\left(1 + \frac{b}{\lambda_{mn}}\right). \tag{3.18}$$

Since $b \in (3, 7)$ and there are no negative eigenvalues of L_T between $\lambda_{10} = -3$ and $\lambda_{21} = -7$, the only negative value of $1 + \frac{b}{\lambda_{mn}}$ occurs at $\lambda_{01} = -3$, which is simple because of our restriction to the subspace $\mathcal{H}$. Therefore,

$$deg(I + B_N, B_\gamma(0) \cap \mathcal{M}_N, 0) = -1$$

and (D2) holds. The proof of the theorem is complete.

Remark 3.5. We note that the theorem holds for other ranges of b. The proof follows exactly; we need only check that, in verifying (D2), we have

$$deg(T_0, B_\gamma(0), 0) = deg(I + bL_T^{-1}, B_\gamma(0), 0) = -1.$$

From (3.17) and (3.18), we see that this amounts to ensuring that $1 + \frac{b}{\lambda_{mn}} < 0$ an odd number of times. For example, if $b \in (11, 15)$, $1 + \frac{b}{\lambda_{mn}} < 0$ for $\lambda_{10} = -3$, $\lambda_{21} = -7$, and $\lambda_{32} = -11$, and the theorem holds. Similarly, if $b \in (15, 19)$, $1 + \frac{b}{\lambda_{mn}} < 0$ for $\lambda_{10}, \lambda_{21}, \lambda_{32}$, and $\lambda_{43} = \lambda_{20} = -15$. (One can verify that there are no other m, n such that $\lambda_{mn} = -15$.)

Remark 3.6. We note that the theorem holds if we change the operator from $L_T u = u_{tt} - u_{xx}$ to $Lu = u_{tt} - au_{xx}$, the domain Ω from $(0, \pi) \times (0, \pi)$ to $(0, \sqrt{a}\pi) \times (0, \pi)$, and adjust the spatial symmetry requirement in the definition of the subspace $\mathcal{H}$ appropriately.

3.3. Continuity of solutions. In this section we prove that, under an additional assumption on the forcing term $h(x, t)$, solutions $u \in \mathcal{H}$ to (3.1) are continuous.

We denote by $\mathcal{H}^m(\Omega)$ or $\mathcal{H}^m$ the Sobolev space $W^{m,2}(\Omega) = \{u | D^\alpha u \in \mathcal{L}^2(\Omega), |\alpha| \leq m\}$, where D^α is a weak derivative. We equip this space with the standard inner product

$$(f, g) = \sum_{|\alpha| \leq m} \int_\Omega D^\alpha f D^\alpha g dA$$

and the norm induced by this inner product.

LEMMA 3.7. *Let the region Ω and the operator L_T^{-1} be as defined above.*

1. *If $w \in \mathcal{L}^2(\Omega)$, then $L_T^{-1} w \in \mathcal{H}^1(\Omega)$.*
2. *If $w \in \mathcal{H}^1(\Omega)$, then $L_T^{-1} w \in \mathcal{H}^2(\Omega)$.*
3. *If $w \in \mathcal{H}^2(\Omega)$, then $w \in \mathcal{C}(\Omega)$.*
3. *If $w \in \mathcal{H}^1(\Omega)$, then $\sin w \in \mathcal{H}^1(\Omega)$.*

Proof. Let $w = \sum_{m,n=0}^{\infty} c_{mn} \phi_{mn} \in \mathcal{L}^2$. It is straightforward to verify that the $\mathcal{L}^2, \mathcal{H}^1$, and $\mathcal{H}^2$ norms of w are given by

$$\begin{aligned} \|w\|_{\mathcal{L}^2}^2 &= \sum c_{mn}^2 < \infty, \\ \|w\|_{\mathcal{H}^1}^2 &= \sum [(2n+1)^2 + (2m)^2] c_{mn}^2, \\ \|w\|_{\mathcal{H}^2}^2 &= \sum [(2n+1)^2 + (2m)^2]^2 c_{mn}^2, \end{aligned}$$

respectively.

1. Let $w = \sum c_{mn} \phi_{mn} \in \mathcal{L}^2$. Then $L_T^{-1} w = \sum \frac{c_{mn}}{\lambda_{mn}} \phi_{mn}$ and

$$\begin{aligned} \|L_T^{-1} w\|_{\mathcal{H}^1}^2 &= \sum [(2n+1)^2 + (2m)^2] \left| \frac{c_{mn}}{\lambda_{mn}} \right|^2 \\ &= \sum \frac{(2n+1)^2 + (2m)^2}{[(2n+1)^2 - (2m)^2]^2} c_{mn}^2 < \infty \end{aligned}$$

since

$$\frac{(2n+1)^2 + (2m)^2}{[(2n+1)^2 - (2m)^2]^2} \leq 1$$

and $\sum c_{mn}^2 < \infty$.

2. This proof is analogous to the proof of (1).
3. See, for example, [1].
4. Let $w \in \mathcal{H}^1$. Then $w, w_t, w_x \in \mathcal{L}^2$; we must show that $\sin w, (\sin w)_t, (\sin w)_x \in \mathcal{L}^2$.

$$\|\sin w\|_{\mathcal{L}^2}^2 = \int_\Omega |\sin w|^2 < \infty$$

since Ω is bounded.

$$\|(\sin w)_t\|_{\mathcal{L}^2}^2 = \int_\Omega |w_t \cos w|^2 \leq \int_\Omega |w_t|^2 < \infty$$

since $w_t \in \mathcal{L}^2$. Similarly, $(\sin w)_x \in \mathcal{L}^2$, and the result follows. □

THEOREM 3.8. *Let $h \in \mathcal{H}^1$, and let $u \in \mathcal{H}$ solve* (3.1). *Then $u \in \mathcal{C}(\Omega)$.*

Proof. The result follows from repeated application of Lemma 3.7. Since $\varepsilon h - b \sin u \in \mathcal{L}^2$, we have $u = L_T^{-1}(\varepsilon h - b \sin u) \in \mathcal{H}^1$. Since $u \in \mathcal{H}^1$, we have $\sin u \in \mathcal{H}^1$, and therefore $h \in \mathcal{H}^1$ implies $\varepsilon h - b \sin u \in \mathcal{H}^1$. It follows then that $u = L_T^{-1}(\varepsilon h - b \sin u) \in \mathcal{H}^2$ and therefore $u \in \mathcal{C}(\Omega)$. □

4. The bifurcation curve of periodic solutions.

In section 3 we considered the forced sine-Gordon equation on a bounded domain, which models the torsional motion of the center span of a suspension bridge, and proved that, under certain assumptions on the physical constants, multiple periodic solutions exist. In this section, we compute periodic solutions to the damped equation and examine their bifurcation properties as the amplitude of the forcing term varies. More specifically, we employ numerical continuation algorithms by which we plot the amplitude of a periodic solution versus the amplitude λ of the external forcing term. We demonstrate that for small λ, multiple periodic solutions to the equation exist. Moreover, we demonstrate that bifurcation from single to multiple periodic solutions occurs for small λ.

Recall from section 2 that the equation that governs the torsional motion along the length of the center span is given by

$$(4.1) \qquad \left\{ \begin{array}{c} \theta_{tt} - \varepsilon_1 \theta_{xx} = -\frac{6K}{m} \cos\theta \sin\theta - \delta\theta_t + h_1(x,t) \\ \theta(0,t) = \theta(L,t) = 0 \end{array} \right\}.$$

For our numerical study of this equation, we must choose the values of the constants $L, m, K, \delta, \varepsilon_1$ and the external forcing term $h_1(x,t)$.

4.1. The choice of physical constants and external forcing.

The length of the span was $L = 1000$ meters [2]; let us normalize the equation so that we can work on the domain $x \in [0,1]$. The rescaled equation is

$$(4.2) \qquad \left\{ \begin{array}{c} \theta_{tt} - \frac{\varepsilon_1}{L^2} \theta_{xx} = -\frac{6K}{m} \cos\theta \sin\theta - \delta\theta_t + h_1(Lx,t) \\ \theta(0,t) = \theta(1,t) = 0 \end{array} \right\}.$$

To determine the physical constants $m, K, \delta, \varepsilon_1$ and the external forcing term $h_1(x,t)$, we rely on [2], [9], and [17]. We choose $m = 2500$ and $\delta = .01$. To determine K, we know from [2] that the main span would deflect about half a meter when loaded with 100 kgs per unit length, so we have $100(9.8) - 2K(.5) = 0$ and we take $K = 1000$. The roadbed of the Tacoma Narrows was extremely flexible, so we choose $\frac{\varepsilon_1}{L^2} = .01$ and observe that this value produces the appropriate flexibility in our numerical solutions.

For a cross section similar to the Tacoma Narrows bridge, wind tunnel experiments indicate that aerodynamic forces should induce approximately sinusoidal oscillations of amplitude three degrees [17], so in (4.2) we choose h_1 to be sinusoidal in time. We take $h_1(x,t) = \lambda \sin(\mu t)\rho(x)$, where $\lambda \in [0, 0.06]$ is chosen to produce the appropriate behavior near equilibrium and the frequency μ is chosen to match the frequency of the oscillations observed at Tacoma Narrows on the day of the collapse. The frequency of the torsional motion was approximately one cycle every 4 or 5 seconds, so we take $\mu \in [1.2, 1.6]$. Thus, (4.2) becomes

$$\left\{ \begin{array}{c} \theta_{tt} - .01\theta_{xx} = -2.4\cos\theta\sin\theta - .01\theta_t + \lambda\sin(\mu t)\rho(x) \\ \theta(0,t) = \theta(1,t) = 0 \end{array} \right\}. \tag{4.3}$$

Remark 4.1. Using $\cos\theta\sin\theta = \frac{1}{2}\sin 2\theta$ and rescaling (4.3), we see that the magnitude of the nonlinear term is $b = 2.4$. Note, however, that Theorem 3.2 does not apply to this problem because of the damping term in (4.3) and the fact that the theorem requires a relationship between the wave speed and the spatial domain that is not satisfied by the physical problem (4.3) (see Remark 3.6). However, as the following experiments demonstrate, the physical problem (4.3) exhibits the multiple solution behavior guaranteed by Theorem 3.2 for the theoretical problem (3.1).

The torsional motion observed on the day of the collapse was, for the most part, one-noded (i.e., no torsional displacement in the middle of the span). Occasionally, the motion changed to no-noded twisting and back again to one-noded. Thus, we take $\rho(x) = 1, \rho(x) = \sin(2\pi x)$, or $\rho(x) = \sin(\pi x)$.

4.2. The numerical results for the forced, damped sine-Gordon equation. In this section, we apply a numerical continuation algorithm to the boundary value problem (4.3) for several different forcing terms:

$$h_1(x,t) = \lambda\sin(\mu t)\rho(x).$$

Numerical continuation algorithms are described in [3], [8], and [15]; we refer the reader to these sources for details.

In each case, we find that if $\mu \in [1.2, 1.5]$, the path of periodic solutions is S-shaped and that bifurcation from single to multiple periodic solutions occurs at a small value of $\lambda = \underline{\lambda}$. Moreover, we observe that $\underline{\lambda}$ decreases as the forcing frequency μ increases.

If μ is greater than the resonant frequency $\hat{\mu}$ of the linearized PDE

$$\theta_{tt} - \varepsilon\theta_{xx} + \delta\theta_t + 2.4\theta = \lambda\sin(\mu t)\rho(x),$$

the amplitude of the periodic solution increases with λ, but bifurcation from single to multiple solutions does not occur. This is consistent with our earlier results for the simpler ODE model [10], [15]. We note that for the space independent, one-noded, and no-noded forcing terms given in section 4.1 above, the resonant frequencies of the linearized PDE are $\hat{\mu} \approx 1.55, \hat{\mu} \approx 1.67$, and $\hat{\mu} \approx 1.58$, respectively.

Forcing independent of x. $h_1(x,t) = \lambda\sin(\mu t)$.

1. Experiment 4.1. $\mu = 1.3, \mu = 1.4, \mu = 1.5$; see Figure 4.1(a). The bifurcation curves are S-shaped, and bifurcation from single to multiple solutions occurs for small $\lambda = \underline{\lambda}$. For example, in Figure 4.1 we see that for $\mu = 1.3$ (the solid curve), a unique small amplitude periodic solution exists for $\lambda < \underline{\lambda} \approx .022$ and $\lambda > \overline{\lambda} \approx .247$ but that three periodic solutions, one of small amplitude and two of large amplitude, exist for $\lambda \in (\underline{\lambda}, \overline{\lambda})$. Whether the small or large

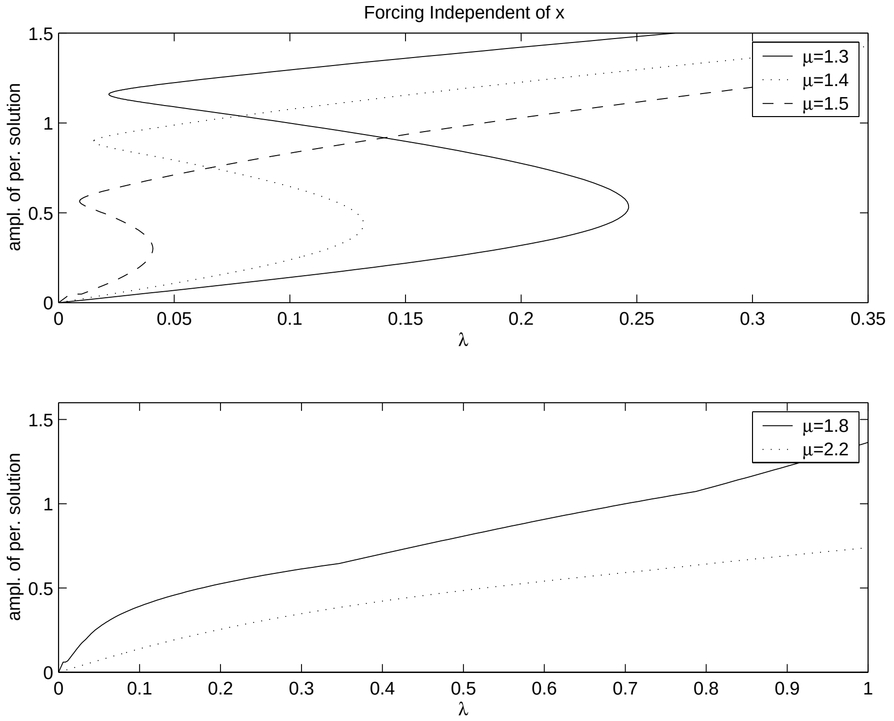


FIG. 4.1. *Experiments* 4.1 *and* 4.2. *At the lower frequencies, there are three periodic solutions under small fixed forcing, one of small amplitude and two of large amplitude. At the higher frequencies, bifurcation from single to multiple periodic solutions does not occur.*

amplitude solution results depends on the initial displacement and velocity of the span. For example, for $\mu = 1.3$, $\lambda \approx 0.047$, under a small initial displacement (approximately 0.003 radians in amplitude), a small amplitude periodic solution results (approximately 0.065 radians). However, under a large initial displacement (approximately 0.720 radians), a large periodic solution results (approximately 1.221 radians). Moreover, we observe that $\underline{\lambda}$, the frequency at which bifurcation from single to multiple periodic solutions occurs, decreases as μ increases.

2. Experiment 4.2. $\mu = 1.8, \mu = 2.2$; see Figure 4.1(b). The amplitude of the periodic solution increases with λ, but bifurcation from single to multiple solutions does not occur. Moreover, we observe that the growth in the amplitude of the periodic solution is slower at the higher frequency. This is consistent with our earlier results for the simpler ODE model [10], [15].

One-noded forcing. $h_1(x,t) = \lambda \sin(\mu t) \sin(2\pi x)$.

1. Experiment 4.3. $\mu = 1.3, \mu = 1.4, \mu = 1.5$; see Figure 4.2(a). Again, the bifurcation curves at these frequencies are S-shaped and our results are consistent with those in Experiment 4.1.
2. Experiment 4.4. $\mu = 1.6, \mu = 1.8, \mu = 2.2$; see Figure 4.2(b). As in Experiment 4.2, we see that the amplitude of the periodic solution grows with λ and that the growth is slower for the higher frequencies. Moreover, for $\mu = 1.6$, we see that multiple periodic solutions exist for a small range of λ. (Recall though that $1.6 < \hat{\mu} \approx 1.67$, the resonant frequency for the linearized PDE.)

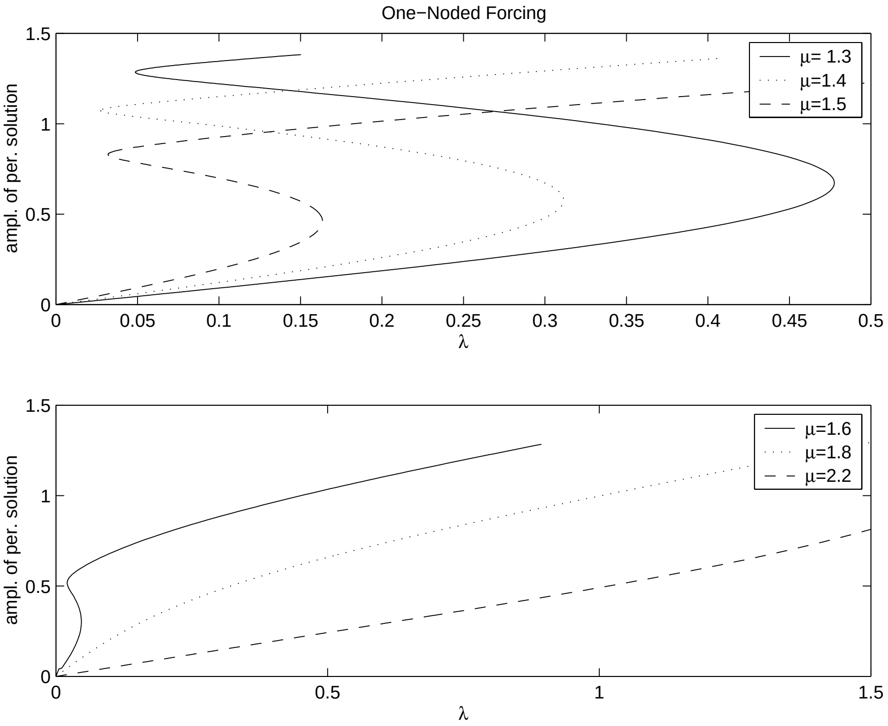


FIG. 4.2. *Experiments* 4.3 *and* 4.4. *As in Experiments* 4.1 *and* 4.2, *at the lower frequencies, bifurcation from single to multiple periodic solutions occurs for small periodic forcing.*

No-noded forcing. $h_1(x,t) = \lambda \sin(\mu t)\sin(\pi x)$.

1. Experiment 4.5. $\mu = 1.3, \mu = 1.4, \mu = 1.5$. Again, the bifurcation curves at these frequencies are S-shaped. As this is consistent with our results in Experiments 4.1 and 4.3, we do not show the bifurcation curves here.
2. Experiment 4.6. $\mu = 1.6, \mu = 1.8, \mu = 2.2$. As in Experiments 4.2 and 4.4, the amplitude of the periodic solution increases with λ, but bifurcation from single to multiple solutions does not occur. Again, the growth in the amplitude of the periodic solution is slower at the higher frequencies. As these results are similar to the earlier experiments, we do not show the figures here.

5. Dynamic response to initial conditions. In section 4, we demonstrated that if $\mu \in [1.2, 1.5]$, under fixed periodic forcing $h_1(x,t) = \lambda \sin(\mu t)\rho(x)$, (4.3) has three periodic solutions: one of small amplitude and two of large amplitude. In this section, we will examine the structural properties of these solutions numerically. More specifically, we will compute solutions to the boundary value problem (4.3) under the initial conditions

$$\begin{aligned} \theta(x,0) &= \xi(x), \\ \theta_t(x,0) &= \eta(x) \end{aligned} \tag{5.1}$$

via finite difference methods. The periodic solution results as the long term solution to the initial value problem; i.e., the span "settles down" to periodic oscillation. As in section 4, we choose $\rho(x) = 1, \rho(x) = \sin(2\pi x)$, or $\rho(x) = \sin(\pi x)$.

Our finite difference scheme is implicit in the linear terms and explicit in the nonlinear terms. We solve the initial value problem (4.3), (5.1) over 400 periods of the forcing term; i.e., for $(x,t) \in [0,1] \times [0, 400\tau]$, where $\tau = \frac{2\pi}{\mu}$. In each experiment we use 520 time steps per period of the forcing term ($\Delta t = \frac{1}{520}\tau$) and we take $\Delta x = .025$.

We define

$$\begin{aligned} a &= \text{amplitude of the initial displacement } \xi(x), \\ a_p &= \text{amplitude of the resulting periodic solution.} \end{aligned}$$

In the experiments that follow we observe that, if $\mu \in [1.2, 1.5]$, under *fixed* periodic forcing $h_1(x,t) = \lambda \sin(\mu t)\rho(x)$, small or large amplitude behavior may result depending only on the initial displacement and velocity of the span. Thus, the effect of a large initial displacement may *not* damp away as in the linear case. Moreover, we find that the amplitude a_p of the periodic response is extremely sensitive to slight changes in the amplitude a of the initial displacement and that a_p does not depend on a in an intuitive way; for example, it does not increase with a. Finally, we observe that the qualitative properties such as amplitude, frequency, and nodal structure of our large amplitude solutions are consistent with the behavior observed at Tacoma Narrows on the day of its collapse.

5.1. The experiments. One-noded forcing and initial conditions. The most prevalent motion observed at Tacoma Narrows was one-noded (no displacement at the center of the span) [2], so let us consider external forcing of the form

$$h_1(x,t) = \lambda \sin(\mu t)\sin(2\pi x)$$

and initial conditions of the form

$$\theta(x,0) = \theta(x,\Delta t) = a\sin(2\pi x).$$

Experiment 5.1. $\lambda = .06, \mu = 1.4$.

- **5.1a.** $\theta(x,0) = \theta(x,\Delta t) = .9\sin(2\pi x)$; see Figure 5.1. Despite the large initial displacement, we see in Figure 5.1 that by periods 390 through 400 of the forcing term, the span has settled down to one-noded, periodic oscillation of small amplitude (approximately .072 radians).
- **5.1b.** $\theta(x,0) = \theta(x,\Delta t) = 1.0\sin(2\pi x)$; see Figure 5.2. We have increased the amplitude a of the initial displacement only slightly from 5.1a, but we see in Figure 5.2 that this small change has a dramatic impact on the motion of the span. As in 5.1a, by periods 390 through 400 of the forcing term, the span has settled down to periodic oscillation. But instead of settling to near equilibrium behavior, as in 5.1a, the amplitude of the oscillation is approximately 1.117 radians. Again, we note that this is close to the amplitude observed at Tacoma Narrows on the day of the collapse [2].
- **5.1c.** See Figure 5.3. Based on our results in Experiments 5.1a and 5.1b, it is tempting to conjecture that the amplitude a_p of the periodic solution increases with the amplitude a of the initial displacement, but this is not the case. Figure 5.3 shows the amplitude a_p of the periodic solution versus the amplitude a of the initial displacement of the span for $a \in [0, 1.7]$. We see in Figure 5.3 that the amplitude of the long term periodic response depends on the amplitude of the initial displacement in an unpredictable way. This is consistent with results for a simple nonlinear ODE model for the vertical

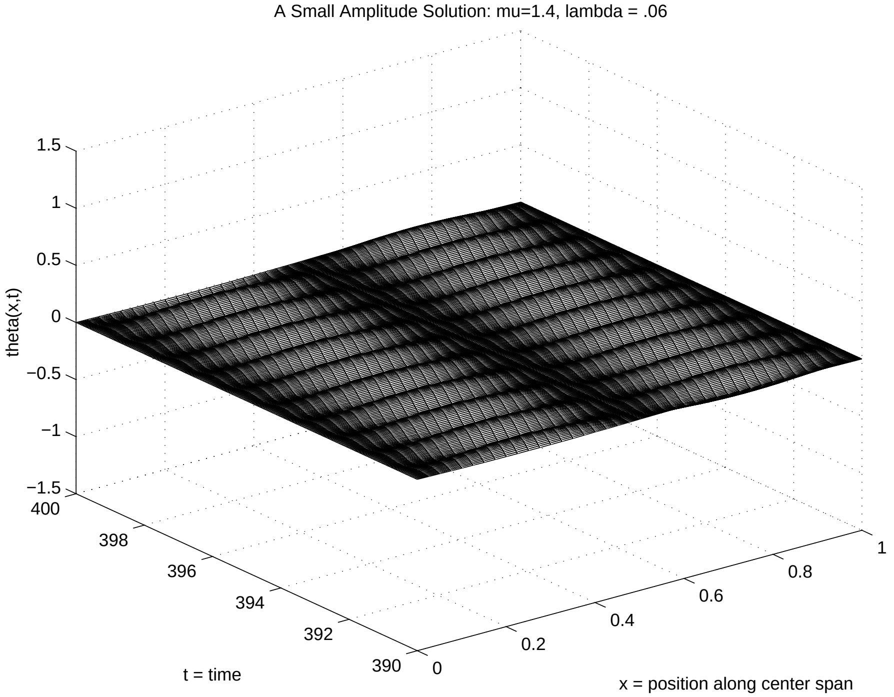


FIG. 5.1. *Experiment* 5.1a.

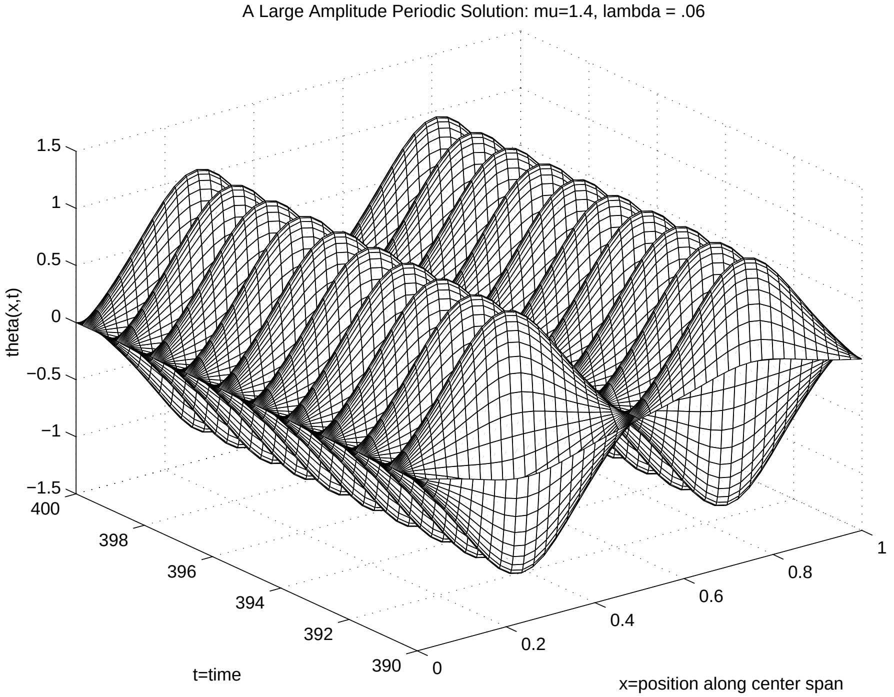


FIG. 5.2. *Experiment* 5.1b.

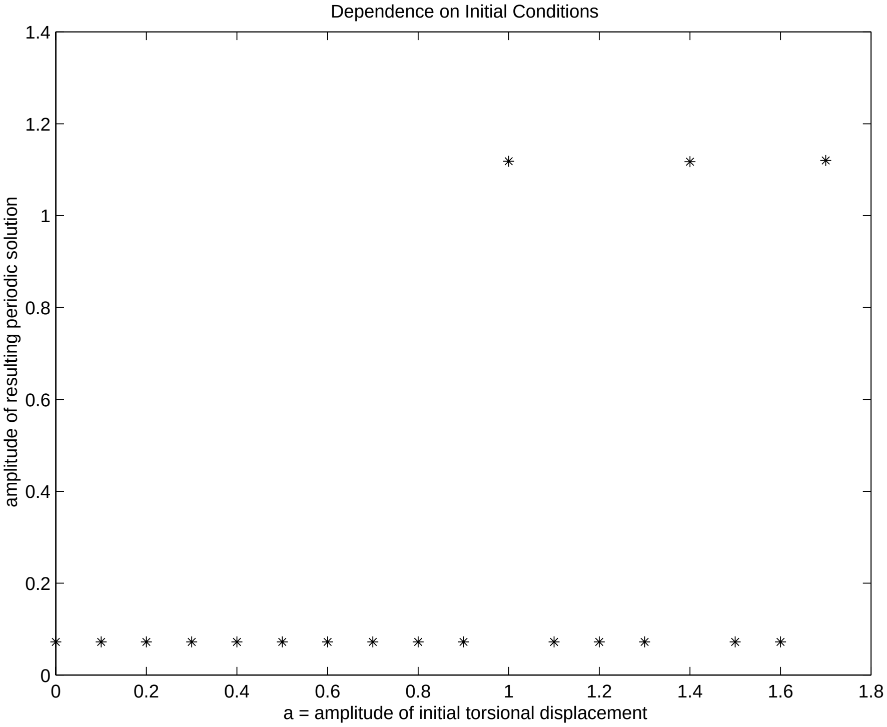


FIG. 5.3. *Experiment* 5.1c.

motion of a suspension bridge [7]. We note that in Figure 5.3, the small solutions correspond to the "bottom branch" of the bifurcation curve in Figure 4.2(a) and the large solutions correspond to the "top branch."

Forcing that depends only on time. We also considered the response of the main span to small, time dependent forcing which is constant along the length of the span, specifically,

$$h_1(x,t) = \lambda \sin(\mu t)$$

and initial conditions of the form

$$\theta(x,0) = \theta(x,\Delta t) = a\sin(2\pi x).$$

Experiment 5.2. $\lambda = .04, \mu = 1.4$. As these results are consistent with those in Experiment 5.1, we do not show the figures; we simply describe the results.

- **5.2a.** $\theta(x,0) = \theta(x,\Delta t) = .5\sin(2\pi x)$. Despite the large initial displacement, by periods 390 through 400 of the forcing term, the span has settled down to no-noded, periodic oscillation of small amplitude (approximately .086 radians).
- **5.2b.** $\theta(x,0) = \theta(x,\Delta t) = .6\sin(2\pi x)$. We have increased the amplitude a of the initial displacement only slightly from 5.2a, but this small change has a dramatic impact on the motion of the span. As in 5.2a, by periods 390 through 400 of the forcing term, the span has settled down to periodic oscillation. But instead of settling to near equilibrium behavior, as in 5.2a, the amplitude of the oscillation is approximately .969 radians. Again, this

is close to the amplitude observed at Tacoma Narrows on the day of the collapse [2].

No-noded forcing and initial conditions. Although the most prevalent mode of torsional oscillation observed at Tacoma Narrows was the one-noded motion described above, occasionally the motion would change to no-noded oscillation [2], so we also studied external forcing of the form

$$h_1(x,t) = \lambda \sin(\mu t) \sin(\pi x)$$

and initial conditions of the form

$$\theta(x,0) = \theta(x,\Delta t) = a \sin(\pi x).$$

As in the previous experiments, small changes in the amplitude of the initial displacement led to dramatic differences in the resulting periodic solution. Indeed, when we *decreased* the amplitude of the initial displacement from 1.2 to 1.1, the amplitude of the resulting periodic solution *increased* from .0248 to 1.171 radians [15].

Solutions that change nodal structure. According to eyewitnesses, the torsional oscillations that preceded the collapse of the Tacoma Narrows were, for the most part, one-noded. Occasionally, the motion would change to no-noded and then back to one-noded [2]. In this experiment, we replicate this phenomenon by a slight perturbation in the forcing term.

Experiment 5.4. $\lambda = .06, \mu = 1.4$; see Figure 5.4. We begin with a large initial displacement

$$\theta(x,t) = \theta(x,\Delta t) = 1.4 \sin(2\pi x)$$

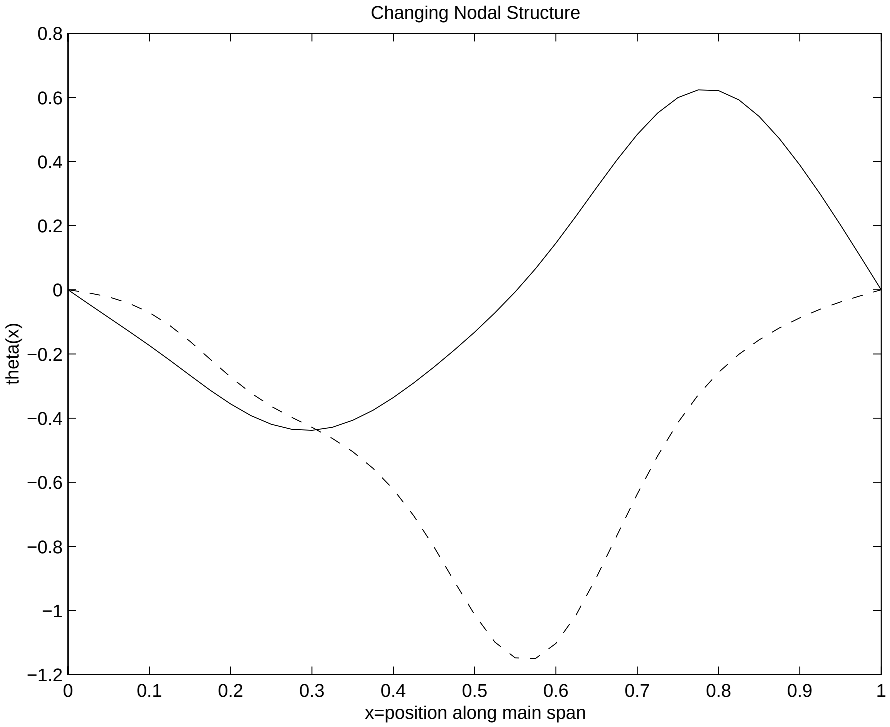


FIG. 5.4. *Experiment* 5.4, *fixed t.*

and apply forcing of the form

$$h_1(x,t) = \lambda \sin(\mu t)[\sin(2\pi x) + .01 \sin(\pi x)].$$

In this case, a complicated motion results. Figure 5.4 shows the angular displacement along the length of the span at two different points in time; the solid curve describes one-noded oscillation while the dashed curve has no nodes.

6. Conclusion and open questions. We have demonstrated theoretically and numerically that the equation that governs the torsional motion of a suspension bridge has multiple periodic solutions; whether small or large amplitude motion results depends on the initial displacement and velocity of the span. Thus, once a large torsional motion starts, it may persist over a long time.

It is natural to ask what might induce such a large initial torsional displacement in a suspension bridge. In studying coupled systems of the form (2.1) numerically, we find that a large vertical motion, in the presence of *tiny* torsional forcing and initial conditions, may induce a rapid transition from vertical to torsional motion [9], [11], [12], [15]. Such a phenomenon was observed at Tacoma Narrows on the day of its collapse [2].

Beyond the results presented here, several interesting questions remain. For example, we proved the existence of multiple periodic solutions to the undamped equation; under appropriate hypotheses on the forcing term, does a similar result hold for the damped equation? Moreover, we proved the existence of at least two periodic solutions, but our numerical results in section 4 suggest that three periodic solutions exist. Can the existence of the third solution be proven?

Acknowledgments. This paper is part of the author's thesis work; she gratefully acknowledges her thesis advisor P. J. McKenna for his guidance. The author also thanks the anonymous referee for the valuable comments and suggestions.

REFERENCES

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

[2] O.H. Amann, T. von Kármán, and G.B. Woodruff, *The Failure of the Tacoma Narrows Bridge,* Federal Works Agency, U.S. National Archives and Records Administration, College Park, MD, 1941.

[3] Y.S. Choi, K.C. Jen, and P.J. McKenna, *The structure of the solution set for periodic oscillations in a suspension bridge model,* IMA J. Appl. Math., 47 (1991), pp. 283–306.

[4] Q. Choi and T. Jung, *The study of a nonlinear suspension bridge equation by a variational reduction method*, Appl. Anal., 50 (1993), pp. 73–92.

[5] P. Drábek, H. Leinfelder, and G. Tajčová, *Coupled string-beam equations as a model of suspension bridges,* Appl. Math., 44 (1999), pp. 97–142.

[6] L.D. Humphreys and P.J. McKenna, *Multiple periodic solutions for a nonlinear suspension bridge equation,* IMA J. Appl. Math., 63 (1999), pp. 37–49.

[7] L.D. Humphreys and R. Shammas, *Finding unpredictable behavior in a simple ordinary differential equation*, College Math. J., 31 (2000), pp. 338–346.

[8] H.B. Keller, *Lectures on Numerical Methods in Bifurcation Problems*, Springer-Verlag, Berlin, 1987.

[9] P.J. McKenna, *Large torsional oscillations in suspension bridges revisited: Fixing an old approximation*, Amer. Math. Monthly, 106 (1999), pp. 1–18.

[10] P.J. McKenna and K.S. Moore, *Multiple periodic solutions to a suspension bridge ordinary differential equation,* Electron. J. Differ. Equ. Conf., 5 (2000) pp. 183–199.

[11] P.J. McKenna and K.S. Moore, *The global structure of periodic solutions of a suspension bridge mechanical model,* IMA J. Appl. Math., submitted.

[12] P.J. McKenna and C. O'Tuama, *Large torsional oscillations in suspension bridges revisited yet again: Vertical forcing creates torsional response,* Amer. Math. Monthly, 108 (2001), pp. 738–745..

[13] P.J. McKenna and W. Walter, *Nonlinear oscillations in a suspension bridge*, Arch. Ration Mech. Anal., 98 (1987), pp. 167–177.

[14] P.J. McKenna and W. Walter, *Traveling waves in a suspension bridge*, SIAM J. Appl. Math., 50 (1990), pp. 703–715.

[15] K.S. Moore, *Large Amplitude Torsional Oscillations in a Nonlinearly Suspended Beam: A Theoretical and Numerical Investigation,* dissertation, University of Connecticut, Storrs, CT, 1999.

[16] B. Moran, *A bridge that didn't collapse,* American Heritage of Invention and Technology, 15 (1999), pp. 10–18.

[17] R.H. Scanlan and J.J. Tomko, *Airfoil and bridge deck flutter derivatives*, Proc. Amer. Soc. Civ. Eng. Eng. Mech. Division, EM6 (1971), pp. 1717–1737.

[18] A. Thorncroft, *London's Millennium Bridge closes,* Financial Times, 12 June 2000.

SIAM J. Math. Anal.
Vol. 33, No. 6, pp. 1430–1454

A FREE BOUNDARY PROBLEM IN DERMAL DRUG DELIVERY[*]

JÁN FILO[†] AND VOLKER PLUSCHKE[‡]

Abstract. In this paper we study a free boundary problem in a multicomponent domain. Our study was motivated by the mathematical modeling of dermal and transdermal drug delivery, where the multilayered skin model was considered. At the interface connecting two components the conservation of the flux and Nernst's distribution law hold and it is supposed that in any component there is a positive minimum concentration at which the diffusion front can proceed. The existence of a solution and uniqueness in special cases are shown.

Key words. free boundary problem, nonlinear diffusion, multicomponent domain, existence, uniqueness

AMS subject classifications. 35K50, 35K55

PII. S0036141001385794

1. Introduction.

Let Ω^i, $i = 1, \dots, k$, $k \geq 2$, be disjoined bounded Lipschitz domains in $\mathbb{R}^N$, $N \geq 1$, with $\bar{\Omega}^i \cap \bar{\Omega}^j = \emptyset$ if $|j-i| > 1$ and

$$\bar{\Gamma}^i \equiv \partial\Omega^i \cap \partial\Omega^{i+1} \qquad \text{for } i = 1, \dots, k-1.$$

Here $\bar{\Gamma}^i$ denotes the closure of a nonempty and relatively open set Γ^i in $\partial\Omega^i$.

In this paper we study the singular parabolic problem

$$\partial_t b^i(u^i) = \Delta u^i \qquad \text{in } \Omega^i \times (0,T) \tag{1.1}$$

with contact conditions on the interfaces

$$\partial_{\nu_i} u^i + \partial_{\nu_{i-1}} u^{i-1} = 0 \quad \text{and} \quad u^i = \psi^{i-1}(u^{i-1}) \qquad \text{on } \Gamma^i \times (0,T) \tag{1.2}$$

and boundary and initial conditions

$$u^i = 0 \qquad \text{on } \Gamma^k \times (0,T), \tag{1.3}$$

$$\partial_{\nu_i} u^i = 0 \qquad \text{on } \mathcal{T}^i \times (0,T), \tag{1.4}$$

$$b^i(u^i) = b_0^i \qquad \text{on } \Omega^i \times \{t=0\}, \tag{1.5}$$

$i = 1, \dots, k$, where Γ^k is a part of $\partial\Omega^k \setminus \Gamma^{k-1}$ that is not excluded to be empty, $\mathcal{T}^i \equiv \partial\Omega^i \setminus \Gamma^i \cup \Gamma^{i-1}$ where we set $\Gamma^0 = \emptyset$. Moreover, ν_i denotes the outer unit normal to $\partial\Omega^i$ and $\partial_\nu u \equiv \nabla u \cdot \nu$, $\partial_t = \partial/\partial t$.

The nonlinearity $b^i(\cdot)$ in (1.1) represents a maximal monotone graph in $\mathbb{R} \times \mathbb{R}$ given by

$$b^i(u) = \beta^i(u) + \Lambda^i \vartheta(u)\,, \qquad \vartheta(u) \equiv \begin{cases} 1, & u > 0, \\ [0,1], & u = 0, \\ 0, & u < 0. \end{cases} \tag{1.6}$$

[*] Received by the editors March 2, 2001; accepted for publication (in revised form) December 21, 2001; published electronically July 9, 2002.
http://www.siam.org/journals/sima/33-6/38579.html

[†] Institute of Applied Mathematics, Comenius University, Mlynská dolina, SK–842 48 Bratislava, Slovakia (Jan.Filo@fmph.uniba.sk). This author was partially supported by VEGA grant 1/4195/97.

[‡] Department of Mathematics and Computer Science, Martin-Luther-University Halle, D–06099 Halle, Germany (pluschke@mathematik.uni-halle.de).

Here $\Lambda^i \geq 0$ are given constants and β^i are continuous monotone increasing functions a.e. differentiable and such that

$$0 < \iota \leq (\beta^i)'(u) < \infty \qquad \text{if } u > 0 \tag{1.7}$$

for a given positive constant ι, $\beta^i(0) = 0$, $i = 1, \dots, k$. $\psi^i : \mathbb{R} \to \mathbb{R}$ are again monotone increasing functions a.e. differentiable and

$$0 < \kappa \leq (\psi^i)' \leq K \tag{1.8}$$

for two positive constants κ, K and $\psi^i(0) = 0$.

As it is seen in (1.6), if $\Lambda^i > 0$, we allow b^i to have jumps. The special cases, however, the standard parabolic equations, that is, $b^i(u) = k^i u$ for $k^i > 0$, and the porous medium type equations (see [2]), e.g.,

$$b^i(u) = |u|^{1/m_i} \operatorname{sign} u \,, \quad m_i > 1 \,,$$

are also included. In both cases, $\Lambda_i = 0$.

As our main goal is to deal with a free boundary problem, the simplest model case a reader can have in mind is the two component system in one space dimension, i.e., $k = 2$, $N = 1$, and we set $\Omega^1 = (0, l_1)$, $\Omega_2 = (l_1, l_2)$. Let us look at a situation in which u^i have a particularly simple structure. Nevertheless, let us in fact formulate a problem for functions $c^i(x,t)$ and $s(t)$ first. Assume a curve

$$\{(x,t) : x = s(t),\ t \in [0,T]\}$$

for monotone increasing function $s : [0,T] \to (0, l_2]$ such that $0 < s(0) < l_1$, $s(t^*) = l_1$ for some $t^* \in (0,T)$, and $s(T) = l_2$ is given. Functions c^i are supposed to satisfy the following:

$$\begin{aligned} &\partial_t c^1 = D^1 c^1_{xx}, && 0 < x < \min\{s(t), l_1\},\ 0 < t \leq T\,, \\ &c^1_x(0,t) = 0, && 0 < t \leq T\,, \\ &c^1(x,0) = c^1_0(x) > \lambda^1, && 0 < x < s(0)\,, \\ &\partial_t c^2 = D^2 c^2_{xx}, && l_1 < x < s(t),\ t^* < t \leq T, \end{aligned} \tag{1.9}$$

and the moving boundary condition

$$c^i(s(t),t) = \lambda^i \,, \qquad \lambda^i \dot{s}(t) = -D^i c^i_x(s(t),t) \tag{1.10}$$

with positive constants D^i, λ^i, where $0 < t < t^*$ if $i = 1$ and $t^* < t < T$ if $i = 2$, respectively, in (1.10). In addition to the Stefan problem in one component we require a nonlinear contact condition on the interface $x = l_1$,

$$\begin{aligned} &c^2(l_1,t) = g\left(c^1(l_1,t)\right), \\ &D^1\, c^1_x(l_1,t) = D^2\, c^2_x(l_1,t), \end{aligned} \qquad t^* < t \leq T\,, \tag{1.11}$$

(for motivation see section 2) where g is a given increasing function such that $g(\lambda^1) = \lambda^2$. Now the following question arises. Having given the data, can we find functions $c^i(x,t), s(t)$ and positive T such that (1.9)–(1.11) are satisfied? In this paper we do not deal with the classical solvability of this problem; instead, we study weak solutions of (1.1)–(1.5). However, problem (1.9)–(1.11) can be rewritten into the first one using the idea of reformulation of the classical Stefan setting into an enthalpy formulation; see, e.g., [13]. So we get the corresponding one-dimensional form of problem (1.1)–(1.5) for $\Lambda^i = \lambda^i$, $u^i = c^i - \lambda^i$ and linear β^i.

Our original aim was to prove the existence and uniqueness for this problem. Nevertheless, uniqueness for the original problem turns out to be difficult and we are able to show uniqueness only for a problem where regular $b^i(\cdot)$ are considered. The paper is organized as follows. It starts in section 2 with derivation of the mathematical model of dermal and transdermal drug delivery, as introduced in [12]. Section 3 contains the existence and uniqueness result for the regularized problem, where the graph $b^i(\cdot)$ is replaced by a monotone Lipschitz continuous regularization $b^i_\varepsilon(\cdot)$ and boundary conditions (1.2) are regularized in the natural way as

$$\begin{aligned} \partial_{\nu_{i-1}} u^{i-1} + n\left(\psi^{i-1}(u^{i-1}) - u^i\right) &= 0, \\ \partial_{\nu_i} u^i + n\left(u^i - \psi^{i-1}(u^{i-1})\right) &= 0, \end{aligned} \qquad \text{on} \quad \Gamma^{i-1} \times (0,T)\,, \tag{1.12}$$

where $0 < \varepsilon \ll 1$ and $1 \ll n$.

Assumptions and the statement of the existence theorem of our original problem (1.1)–(1.5) are given in section 4. Finally in section 5 the uniqueness result for the related nonlinear diffusion problem is discussed.

In recent years questions like global solvability, uniqueness, and qualitative behavior of solutions for nonlinear parabolic problems including the Stefan problem have attracted considerable interest. There is also a vast amount of literature that covers our result in case of a single domain; see, e.g., [1], [13], [4], [9], and references therein for existence results and, e.g., [1], [4], [6], [8] for uniqueness. Should we restrict ourselves to parabolic problems on multicomponent domains with a nonlinear contact condition on the interface, that in our case is nonlinear Nernst's law (1.2), then we find that these have not been studied to that extent. See [11] or [10], where linear parabolic equations with contact conditions (1.2) and conditions of the type (1.12), respectively, for the special case $\psi^i(u) = u$ are investigated. But, as far as we know, the nonlinear parabolic problem on a multicomponent domain with jump conditions on the interface between two components has not been studied in connection with the free boundary problem. Most methods that we apply are known; nevertheless, their applications to the difficult problem that comes from the application [12] seem to be not straightforward. To prove existence of a weak solution to our free boundary problem we follow the ideas of Alt and Luckhaus [1]. It is more appropriate for our contact condition than the methods of Meirmanov [13], where the equation is written in terms of $U = b(u)$ and all derivatives are given to the test function. However, to derive a priori estimates for the regularized problem it is not possible to test with u^i on all components like in [10], or in case of one component [1], since the integrals arising from boundary condition (1.12) do not provide nonnegative items. Therefore, in sections 3 and 5 we use the dual problem to derive L^1-estimates, and in section 4 we adopt some ideas from Carrillo [4] to test with nonlinear functions in terms of u.

We finish this section by introducing some notation. In what follows, if necessary, we shall consider any function $u(x)$ defined almost everywhere on some open set $\Omega \subset \mathbb{R}^N$, $u \in H^1(\Omega)$ to be extended outside of Ω (and denoted again by u) such that $\|u\|_{H^1(\mathbb{R}^N)} \le C\|u\|_{H^1(\Omega)}$ with C independent of u. Due to the result of Calderon–Stein (see, e.g., [14]) this is possible if $\partial\Omega$ is Lipschitz. By $\langle \cdot\,, \cdot \rangle$ we denote the duality between $H^1(\Omega)$ and $H^{-1}(\Omega)$. The function spaces we use are rather familiar and we omit the definition (see, e.g., [14]). To keep the notation short we set $V_i \equiv H^1(\Omega^i)$ and $V = V_1 \times \cdots \times V_k$. $\tilde V \subset V$ denotes the subspace $\tilde V = \{v = (v^1, \ldots, v^k) \in V :$

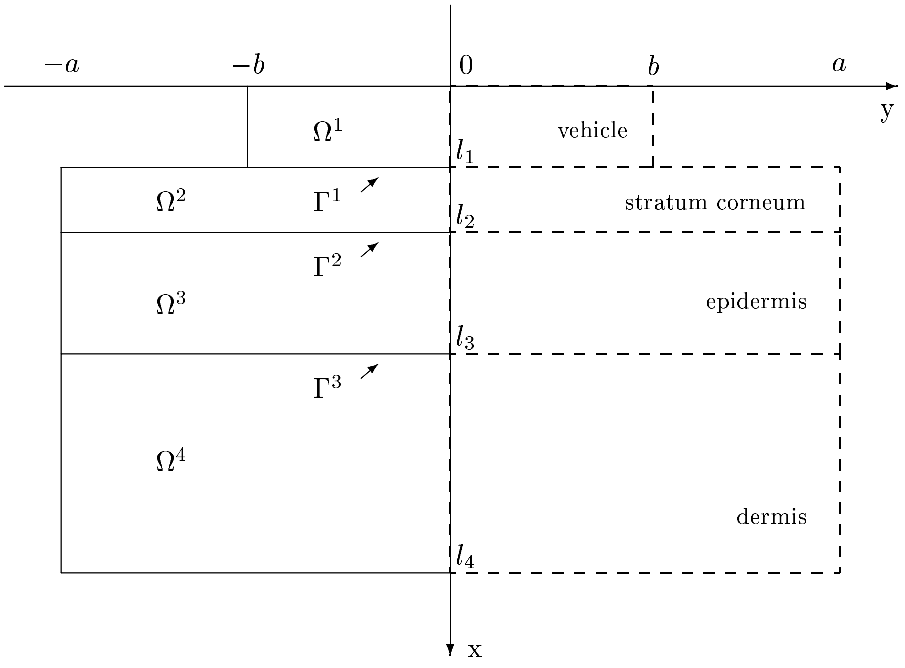


FIG. 1. *The multilayer skin model.* $l_2, \dots, l_4$ *are chosen in accordance with the thickness of the skin layers;* l_1 *represents the application thickness and* $2b$ *its length.* a *either marks the position of the real isolating boundaries in ex vivo experiments or is chosen to be large enough such that the fluxes across* $y = \pm a$ *are negligible.*

$v^{i+1} = v^i$ a.e. on Γ_i for $i = 1, \dots, k-1$, $v^k = 0$ a.e. on $\Gamma_k\}$. Finally, if S is any set in $\mathbb{R}^N$, we write $S_T \equiv S \times (0,T)$ and $S_0 \equiv S \times \{t = 0\}$.

2. Motivation. Our study was motivated by the mathematical modeling of dermal and transdermal drug delivery, where the multilayered skin model in $\mathbb{R}^2$ was considered; see [12]. In this model the whole diffusion area is divided into four open subdomains Ω^i representing the vehicle and the layers of skin, connected by the interfaces Γ^i (see Figure 1), $i = 1, \dots, 4$.

The drug concentration in the subdomains Ω^i is denoted by $c^i = c^i(x,y,t)$. The penetration is described by the nonlinear diffusion equation

$$\partial_t c^i = \operatorname{div}\left(D^i(c^i)\,\nabla c^i\right) \tag{2.1}$$

which relates the change of the concentration c^i in time t to the substance flux $-D^i\,\nabla c^i$ with diffusivity D^i due to Fick's laws. At the interfaces Γ^i connecting two subdomains Ω^i and Ω^{i+1}, the conservation of flux and Nernst's distribution law must hold. These relations are given by

$$D^i\,\partial_x c^i = D^{i+1}\partial_x c^{i+1} \tag{2.2}$$

and

$$c^{i+1} = K^{i+1}\,c^i \tag{2.3}$$

with the partition coefficient being $K^{i+1} > 1$.

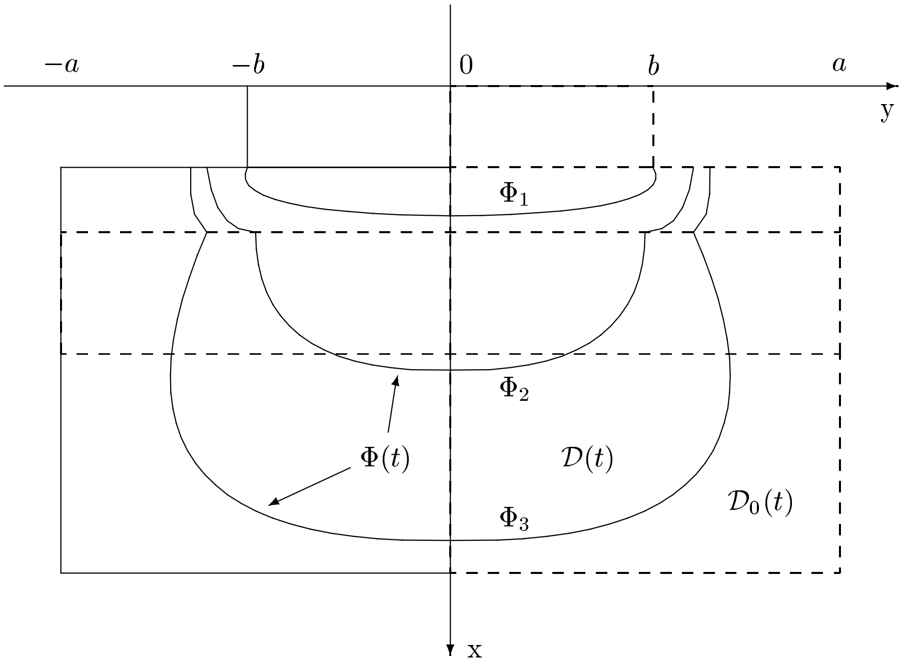


FIG. 2. *Time-dependent penetration depths. The penetration boundaries* Φ_1, Φ_2, Φ_3 *at three time levels* $0 < t_1 < t_2 < t_3$ *are shown; at the corresponding time the concentration is zero outside these boundaries.*

Since penetrants can move into and through human skin only at a finite penetration velocity, it is supposed that a regular curve $\Phi(t)$ in $\Omega \equiv \bigcup(\Omega^i \cup \Gamma^i)$ appears separating Ω into two time-dependent subareas $\mathcal{D}_0(t)$ with zero and $\mathcal{D}(t)$ with nonzero concentration, as illustrated in Figure 2.

Until $\Phi(t)$ reaches the boundary of the Ω^4, the whole mass of drug is inside $\mathcal{D}(t)$. We require the concentration of the drug to have a fixed positive value λ at the boundary $\Phi(t)$ of $\mathcal{D}(t)$. In this context λ describes the minimum concentration at which any diffusion front can proceed. So we set

$$c^i(x,y,t) = \lambda^i \quad \text{for} \quad (x,y) \in \Phi(t) \cap \Omega^i \,. \tag{2.4}$$

Moreover, on $\Phi(t)$ we assume that the penetration velocity of the drug is proportional to the concentration gradient at $\Phi(t)$; i.e.,

$$\lambda^i \, V_\nu = -D^i \, \nabla c^i \cdot \nu, \tag{2.5}$$

where V_ν means the normal velocity of the moving surface $\Phi(t)$ and ν represents its outer normal. To complete the problem formulation it is necessary to prescribe the initial position of the free boundary and the initial concentration distribution.

With the usual Kirchhoff transformation

$$u^i \;=\; \int_{\lambda^i}^{c^i} D^i(z)\, dz$$

denoting

$$c^i - \lambda^i = \beta^i(u^i);$$

the relations (2.1)–(2.5) above yield

$$\partial_t \beta^i(u^i) = \Delta u^i$$

in $\big(\mathcal{D}(t) \cap \Omega^i\big) \times (0,T)$,

$$\partial_x u^i = \partial_x u^{i+1} \quad \text{and} \quad \beta^{i+1}(u^{i+1}) = K^{i+1}\left(\beta^i(u^i) + \lambda^i\right) - \lambda^{i+1} \tag{2.6}$$

on $(\mathcal{D}(t) \cap \Gamma_i) \times (0,T)$,

$$\beta^i(u^i) = 0 \qquad \text{and} \qquad \lambda^i V_\nu = -\nabla u^i \cdot \nu$$

on $\big(\Phi(t) \cap \Omega^i\big) \times (0,T)$, and

$$\beta^i(u^i) = \beta_0^i$$

on $\mathcal{D}(0) \cap \Omega^i$, $i = 1, \dots, 4$.

We shall consider the weak or enthalpy formulation of the above problem, where all references to the free (unknown) boundary $\Phi(t)$ disappear and a problem on a given fixed domain Ω_T will be considered (see, e.g., [13] and references therein). Nernst's distribution law (2.6) is, however, well defined only on $\mathcal{D} \cap \Gamma^i$; therefore we first modify this law setting

$$\beta^{i+1}(u^{i+1}) = K^{i+1}\beta^i(u^i) + L^{i+1}\,\chi(\beta^i(u^i)), \qquad L^{i+1} \equiv K^{i+1}\lambda^i - \lambda^{i+1} > 0\,,$$

$$\chi(\sigma) = \begin{cases} 1, & \sigma > 0, \\ 0, & \sigma \leq 0, \end{cases} \tag{2.7}$$

which is fulfilled on $\mathcal{D}_0 \cap \Gamma^i$, too. Then, due to the noncontinuity of χ_0, we propose the following regularization:

$$\beta^{i+1}(u^{i+1}) = K^{i+1}\beta^i(u^i) + L^{i+1}\,\chi_\delta(\beta^i(u^i)),$$

where

$$\chi_\delta(\sigma) = \begin{cases} 0, & \sigma \leq 0, \\ \sigma/\delta, & 0 \leq \sigma \leq \delta, \\ 1, & 0 < \delta \leq \sigma. \end{cases} \tag{2.8}$$

Here δ is a small parameter, as far as we know, without a physical meaning and we keep it fixed. In our opinion, however, this regularization may be interpreted as a transition stage for small drug concentrations until Nernst's equilibrium is reached.

In this way we have arrived at problem (1.1)–(1.5).

3. Auxiliary problems.

Assume for this section that for given $0 < \varepsilon \ll 1$

$$b_\varepsilon^i(u) = \beta_\varepsilon^i(u) + \Lambda^i \chi_\varepsilon(u), \tag{3.1}$$

$\beta^i_\varepsilon(\cdot)$ are monotone increasing differentiable functions, and

$$0 < \varepsilon \le (\beta^i_\varepsilon)' \le K_\varepsilon \tag{3.2}$$

for a positive constant K_ε, $\beta^i_\varepsilon(0) = 0$, $i = 1, \dots, k$, χ_ε given by (2.8). Moreover, let

$$\beta^i_\varepsilon \longrightarrow \beta^i \qquad \text{as } \ \varepsilon \to 0 \tag{3.3}$$

uniformly on compact subsets of $\mathbb{R}$ and assume that $0 < \iota \le (\beta^i_\varepsilon)' \le K_\varepsilon$ on $[0, \infty)$ (cf. (1.7)). Throughout this section we suppose

$$u^i_{0\varepsilon} \in H^1(\Omega^i) \cap L^\infty(\Omega^i)\,, \quad u^i_{0\varepsilon} \ge 0\,, \tag{3.4}$$

$i = 1, \dots, k$, and

$$b^i_\varepsilon(u^i_{0\varepsilon}) \longrightarrow b^i_0 \qquad \text{in } \ L^1(\Omega^i) \tag{3.5}$$

as $\varepsilon \to 0$.

Given a positive integer n, now consider the system

$$\begin{aligned}
&\partial_t b^i_\varepsilon(u^i_\varepsilon) = \Delta u^i_\varepsilon && \text{in } \ \Omega^i_T,\\
&\partial_{\nu_i} u^i_\varepsilon + n\left(u^i_\varepsilon - \psi^{i-1}(u^{i-1}_\varepsilon)\right) = 0 && \text{on } \ \Gamma^{i-1}_T,\\
&\partial_{\nu_i} u^i_\varepsilon + n(\psi^i(u^i_\varepsilon) - u^{i+1}_\varepsilon) = 0 && \text{on } \ \Gamma^i_T,\\
&\partial_{\nu_i} u^i_\varepsilon = 0 && \text{on } \ \mathcal{T}^i_T,\\
&u^i_\varepsilon = u^i_{0\varepsilon} && \text{on } \ \Omega^i_0
\end{aligned} \tag{3.6}$$

for $i = 1, \dots, k$, where ψ^i satisfies (1.8), $\Gamma^0 = \emptyset$, $u^{k+1} \equiv 0$, and $\psi^k(u) \equiv u$.

Note that our intention in the next section is to send $\varepsilon \to 0$ and $n \to \infty$. But the results of this section might be of interest on their own. Theorem 3.2 extends the existence result of Kačur and Van Keer [10] to the nonlinear problem. Moreover, we prove a maximum principle and continuous dependence of the solution on the initial data for problem (3.6). The same constant n in the contact condition on each interface Γ^i is not essential; it may be replaced by positive constants k^i. Throughout the rest of this section, therefore, ε and n are fixed and we shall not indicate the dependence of the appeared quantities on them.

Definition 3.1. *We say that $u \equiv (u^1, \dots, u^k) \in L^2(0, T; V)$ is a weak solution of the system (3.6) provided $b^i(u^i) \in L^2(\Omega^i_T)$, $\partial_t b^i(u^i) \in L^2(0, T; H^{-1}(\Omega^i))$; that is,*

$$\int_0^T \langle \partial_t b^i(u^i), v^i \rangle dt = -\int_{\Omega^i_T} (b^i(u^i) - b^i(u^i_0))\, \partial_t v^i \, dxdt \tag{3.7}$$

for any $v = (v^1, \dots, v^k) \in L^2(0, T; V)$ with $\partial_t v^i \in L^2(\Omega^i_T)$, $v^i(T) = 0$, and if the following identity is fulfilled:

$$\begin{aligned}
\sum_{i=1}^k \left(\int_0^T \langle \partial_t b^i(u^i), \phi^i \rangle dt \ + \ \int_0^T \int_{\Omega^i} \nabla u^i \nabla \phi^i \, dxdt \right)&\\
+ \sum_{i=1}^k n \int_0^T \int_{\Gamma^i} \left(u^{i+1} - \psi^i(u^i)\right) \, \left(\phi^{i+1} - \phi^i\right) \, d\sigma dt = 0&
\end{aligned} \tag{3.8}$$

for any $\phi = (\phi^1, \dots, \phi^k) \in L^2(0, T; V)$.

A subsolution (supersolution) is defined by (3.8) *with equality replaced by* $\leq$ ($\geq$) *and* $\phi \geq 0$, *i.e.,* $\phi^i \geq 0$ *on* Ω^i *for each* $i \in \{1, \dots, k\}$.

THEOREM 3.2. *A weak solution of problem* (3.6) *exists if conditions* (3.1), (3.2), *and* (3.4) *are fulfilled. Moreover,*

$$u^i \geq 0 \qquad \text{a.e. on} \quad \Omega^i_T. \tag{3.9}$$

In addition, if

$$\left\| u_0^1 \right\|_{L^\infty(\Omega^1)} \leq c^1 \qquad \text{and} \qquad \left\| u_0^i \right\|_{L^\infty(\Omega^i)} \leq c^i \ , \qquad c^i \equiv \psi^{i-1}(c^{i-1})$$

for $i = 2, \dots, k$, *then*

$$\left\| u^i \right\|_{L^\infty(\Omega^i_T)} \leq c^i \tag{3.10}$$

for each $i = 1, \dots, k$.

Let us postpone the proof of Theorem 3.2 to the end of this section.

THEOREM 3.3. (i) *Let* u *and* v *be weak solutions of the problem* (3.6) *with initial functions* u_0 *and* v_0, *respectively. Then for almost all* $t \in [0, T]$,

$$\sum_{i=1}^k \int_{\Omega^i} \left| b^i(u^i) - b^i(v^i) \right| (x,t)\, dx \ \leq \ \sum_{i=1}^k \int_{\Omega^i} \left| b^i(u_0^i) - b^i(v_0^i) \right| (x)\, dx. \tag{3.11}$$

(ii) *Let* u *be a subsolution and* v *a supersolution of the problem* (3.6) *with initial data* u_0 *and* v_0, *respectively. Then if* $u_0 \leq v_0$, *i.e.,* $u_0^i \leq v_0^i$ *a.e. on* Ω^i *for any* $i \in \{1, \dots, k\}$, *it follows that*

$$u^i(x,t) \leq v^i(x,t)$$

a.e. on Ω^i_T *for any* $i \in \{1, \dots, k\}$.

The idea of the proof is as follows. Supposing that we have two solutions u and v, we write down the weak formulation for their difference. Some integrals of the obtained relation may be interpreted as a weak formulation of a suitable boundary value problem for the test function ϕ, the so called dual problem, which is tested by $u - v$. If we insert a solution of that dual problem as a test function, some integrals disappear and we are able to derive L^1-estimates if we know that the solution ϕ of the dual problem is bounded. This method was used, e.g., by [3], [6], [8], and many others.

In our case we have to find such conditions for the dual problem on the interfaces Γ^i that the integrals on the right-hand side of (3.12) may be nullified and ϕ fulfills a maximum principle (cf. (3.13)). However, since the coefficients of the dual problem are not smooth enough we have to regularize it and we have to be very careful to go to the limit.

This idea is used again to prove the boundedness of solution in Theorem 3.2 and the uniqueness result in section 5.

Proof of Theorem 3.3. We first prove the comparison principle (ii). In order to derive the dual problem let $0 < t \leq T$ be fixed. Take $\phi^i(x, \tau) = \chi_{[0,t]}(\tau)\varphi^i(x, t - \tau)$, where the choice of appropriate functions $\varphi^i = \varphi^i(x, s)$ will be determined later and

χ approaches the characteristic function of the interval $[0,t]$. Recalling the definition of u, v, one can see that

$$\begin{aligned}
&\int_{\Omega^i} \left(b^i(u^i) - b^i(v^i)\right)(x,t)\, \varphi^i(x,0)\, dx \\
&\qquad + \int_0^t \int_{\Omega^i} \big(A^i(x,\tau)\, \partial_s \varphi^i(x,t-\tau)\, (u^i - v^i)(x,\tau) \\
&\qquad\qquad + \nabla \varphi^i(x,t-\tau)\, \nabla(u^i(x,\tau) - v^i(x,\tau))\big)\, dx d\tau \\
&\qquad + n \int_0^t \int_{\Gamma^{i-1}} (u^i - v^i)(x,\tau)\, \varphi^i(x,t-\tau)\, d\sigma d\tau \\
&\qquad + n \int_0^t \int_{\Gamma^i} c^i(x,\tau)\, (u^i - v^i)(x,\tau)\, \varphi^i(x,t-\tau)\, d\sigma d\tau \\
&\quad \le \int_{\Omega^i} \left(b^i(u^i_0) - b^i(v^i_0)\right) \varphi^i(x,t)\, dx \\
&\qquad + n \int_0^t \int_{\Gamma^{i-1}} \left(\psi^{i-1}(u^{i-1}) - \psi^{i-1}(v^{i-1})\right)(x,\tau)\, \varphi^i(x,t-\tau)\, d\sigma d\tau \\
&\qquad + n \int_0^t \int_{\Gamma^i} (u^{i+1} - v^{i+1})(x,\tau)\, \varphi^i(x,t-\tau)\, d\sigma d\tau,
\end{aligned} \tag{3.12}$$

where

$$A^i(x,\tau) \equiv \frac{b^i(u^i) - b^i(v^i)}{u^i - v^i}(x,\tau) \qquad \text{and} \qquad c^i(x,\tau) \equiv \frac{\psi^i(u^i) - \psi^i(v^i)}{u^i - v^i}(x,\tau).$$

From the assumed hypotheses one can see that $\varepsilon \le A^i \le K_\varepsilon$ and $0 < \kappa \le c^i \le K$. Formally, if we insert $\varphi = \varphi(x, t-\tau)$ into (3.12), $0 \le \varphi^i \le 1$ such that $\varphi^i(x,s)$, $s \in [0,t]$, satisfies

$$\begin{aligned}
&A^i(x,t-s)\, \partial_s \varphi^i - \Delta \varphi^i = 0 && \text{in } \Omega^i_t, \\
&\partial_{\nu_i} \varphi^i + n\left(\varphi^i - \chi((u^i - v^i)(x,t-s))\right) = 0 && \text{on } \Gamma^{i-1}_t, \\
&\partial_{\nu_i} \varphi^i + n\, c^i(x,t-s)\left(\varphi^i - \chi((u^i - v^i)(x,t-s))\right) = 0 && \text{on } \Gamma^i_t, \\
&\partial_{\nu_i} \varphi^i = 0 && \text{on } \mathcal{T}^i_t, \\
&\varphi^i = \chi(u^i(x,t) - v^i(x,t)) && \text{on } \Omega^i \times \{s = 0\},
\end{aligned} \tag{3.13}$$

χ being given by (2.7). We get

$$\begin{aligned}
&\int_{\Omega^i} \left(b^i(u^i) - b^i(v^i)\right)_+ (x,t)\, dx \;+\; n \int_0^t \int_{\Gamma^{i-1}} (u^i - v^i)_+(x,\tau)\, d\sigma d\tau \\
&\qquad + n \int_0^t \int_{\Gamma^i} \left(\psi^i(u^i) - \psi^i(v^i)\right)_+ (x,\tau)\, d\sigma d\tau \\
&\quad \le \int_{\Omega^i} \left(b^i(u^i_0) - b^i(v^i_0)\right)_+ (x)\, dx \\
&\qquad + n \int_0^t \int_{\Gamma^{i-1}} \left(\psi^{i-1}(u^{i-1}) - \psi^{i-1}(v^{i-1})\right)_+ (x,\tau)\, d\sigma d\tau \\
&\qquad + n \int_0^t \int_{\Gamma^i} \left(u^{i+1} - v^{i+1}\right)_+ (x,\tau)\, d\sigma d\tau,
\end{aligned} \tag{3.14}$$

where

$$w_+ \equiv \max\{0, w\}.$$

If we add (3.14) through $i = 1, \dots, k$, we arrive at

$$\sum_{i=1}^{k} \int_{\Omega^i} \left(b^i(u^i) - b^i(v^i)\right)_+ (x,t)\, dx \le \sum_{i=1}^{k} \int_{\Omega^i} \left(b^i(u_0^i) - b^i(v_0^i)\right)_+ (x)\, dx, \tag{3.15}$$

recalling that we have put $\Gamma^0 = \emptyset$, $u^{k+1} = v^{k+1} = 0$, and $\psi^k(u) \equiv u$.

Nevertheless, coefficients in (3.13) are only bounded functions and we cannot, therefore, expect the required regularity of solutions to (3.13). Therefore, instead of (3.13) we shall consider the problem

$$\begin{aligned}
&A^i_\rho(x,t-s)\,\partial_s\varphi^i - \Delta\varphi^i = 0 && \text{in } \ \Omega^i_t,\\
&\partial_{\nu_i}\varphi^i + n\left(\varphi^i - \omega^i_\delta(x,t-s)\right) = 0 && \text{on } \ \Gamma^{i-1}_t,\\
&\partial_{\nu_i}\varphi^i + n\; c^i_\delta(x,t-s)\left(\varphi^i - \omega^i_\delta(x,t-s)\right) = 0 && \text{on } \ \Gamma^i_t,\\
&\partial_{\nu_i}\varphi^i = 0 && \text{on } \ \mathcal{T}^i_t,\\
&\varphi^i = \varpi^i_\eta(x) && \text{on } \ \Omega^i \times \{s=0\},
\end{aligned} \tag{3.16}$$

where

$$A^i_\rho(x,t-s) \equiv (R_\rho * A^i)(x,t-s), \quad \omega^i_\delta(x,t-s) \equiv \left(R_\delta * \chi(u^i - v^i)\right)(x,t-s),$$
$$c^i_\delta(x,t-s) \equiv \left(R_\delta * c^i\right)(x,t-s),$$

R_ρ, R_δ are the standard mollifiers in the t-variable, and

$$\varpi^i_\eta(x) \equiv \left(R_\eta * \chi(u^i(\cdot,t) - v^i(\cdot,t))\right)(x),$$

R_η is the standard mollifier in the x-variable.

(3.16) is a uniformly parabolic problem on each domain Ω^i_t with sufficiently smooth data and we can apply the classical results of [11] to get a unique weak solution φ^i such that

$$\varphi^i \equiv \varphi^i_{\rho\delta\eta} \in L^\infty(0,t;H^1(\Omega^i)) \cap H^1(0,t;L^2(\Omega^i))$$

and

$$0 \le \varphi^i \le 1 \qquad \text{a.e. on } \ \Omega^i_t. \tag{3.17}$$

To prove (3.17), note that the weak formulation of (3.16) easily gives

$$\int_0^t \int_{\Omega^i} \left(A^i_\rho \partial_s(\varphi^i - 1)\,\xi + \nabla(\varphi^i - 1)\nabla\xi\right)\, dxds \;+\; n\int_0^t \int_{\Gamma^{i-1}} (\varphi^i - 1)\xi\, d\sigma ds$$
$$+\, n\int_0^t \int_{\Gamma^i} c^i_\delta\,(\varphi^i - 1)\,\xi\, d\sigma ds \;\le\; 0$$

for any $\xi \in L^2(0,T;H^1(\Omega^i))$, $\xi \ge 0$. If we insert $\xi = \chi_\varepsilon(\varphi^i - 1)$ and if $\varepsilon \to 0$, we arrive at

$$\int_0^t \int_{\Omega^i} \partial_s\left(A^i_\rho\,(\varphi^i - 1)_+\right)\, dxds \;\le\; \int_0^t \int_{\Omega^i} \left|\partial_s A^i_\rho\right| (\varphi^i - 1)_+\, dxds.$$

This, due to Gronwall's lemma, yields

$$\int_{\Omega^i} (\varphi^i - 1)_+(x,t)\, dx \le 0, \qquad \text{i.e., } (3.17)_2.$$

Analogously one gets

$$\int_{\Omega^i} (-\varphi^i)_+(x,t)\, dx \le 0, \qquad \text{i.e., } (3.17)_1.$$

Moreover, it is easy to see (cf., e.g., [11]) that

$$\max_{0\le s\le t} \int_{\Omega^i} \left|\nabla\varphi^i\right|^2 (x,s)\, dx + \int_0^t \int_{\Omega^i} A^i_\rho(x,t-s) \left|\partial_s\varphi^i\right|^2 dxds \le C, \tag{3.18}$$

where the positive constant $C = C(\eta,\delta)$ does not depend on ρ.

Inserting $\varphi^i_{\rho\delta\eta}$ into (3.12) now we get

$$\begin{aligned}
&\int_{\Omega^i} \left(b^i(u^i) - b^i(v^i)\right)(x,t)\, \varpi^i_\eta(x)\, dx \\
&\qquad + \int_0^t \int_{\Omega^i} \left(A^i - A^i_\rho\right)(x,\tau)\, \partial_s\varphi^i_{\rho\delta\eta}(x,t-\tau)\, (u^i - v^i)(x,\tau)\, dxd\tau \\
&\qquad + n\int_0^t \int_{\Gamma^{i-1}} (u^i - v^i)(x,\tau)\, \omega^i_\delta(x,\tau)\, d\sigma d\tau \\
&\qquad + n\int_0^t \int_{\Gamma^i} (c^i - c^i_\delta)(x,\tau)\, \varphi^i_{\rho\delta\eta}(x,t-\tau)\, (u^i - v^i)(x,\tau)\, d\sigma d\tau \\
&\qquad + n\int_0^t \int_{\Gamma^i} c^i_\delta(x,\tau)\, \omega^i_\delta(x,t-\tau)\, (u^i - v^i)(x,\tau)\, d\sigma d\tau \\
&\quad \le \int_{\Omega^i} \left(b^i(u^i_0) - b^i(v^i_0)\right)_+ dx \\
&\qquad + n\int_0^t \int_{\Gamma^{i-1}} \left(\psi^{i-1}(u^{i-1}) - \psi^{i-1}(v^{i-1})\right)_+ (x,\tau)\, d\sigma d\tau \\
&\qquad + n\int_0^t \int_{\Gamma^i} \left(u^{i+1} - v^{i+1}\right)_+ (x,\tau)\, d\sigma d\tau.
\end{aligned}$$

First we go with $\rho \to 0$ so that the integral with $\partial_s\varphi^i_{\rho\delta\eta}$ disappears. Afterwards we let $\delta \to 0$ and then $\eta \to 0$. Hence we arrive at (3.15) and the desired comparison principle follows.

To prove assertion (i), note that u, v are both sub- and supersolutions, and therefore analogously one gets

$$\sum_{i=1}^k \int_{\Omega^i} \left(b^i(v^i) - b^i(u^i)\right)_+ (x,t)\, dx \le \sum_{i=1}^k \int_{\Omega^i} \left(b^i(v^i_0) - b^i(u^i_0)\right)_+ (x)\, dx.$$

This proves (i). □

Proof of Theorem 3.1. 1. We intend to build a weak solution of system (3.6) by first constructing solutions of certain simpler approximations to (3.6) and then passing to limits. More precisely, now fix a positive integer ℓ. For given

$$u_{\ell-1} = (u^1_{\ell-1}, \dots, u^k_{\ell-1})\,, \quad u^i_0(x,t) \equiv u^i_0(x),$$

we will look for a function $u_\ell \equiv (u_\ell^1, \dots, u_\ell^k)$ so that

$$\begin{aligned}
&\partial_t b^i(u_\ell^i) = \Delta u_\ell^i && \text{in } \Omega_T^i,\\
&\partial_{\nu_i} u_\ell^i + n\left(u_\ell^i - \psi^{i-1}(u_\ell^{i-1})\right) = 0 && \text{on } \Gamma_T^{i-1},\\
&\partial_{\nu_i} u_\ell^i + n\left(\psi^i(u_\ell^i) - u_{\ell-1}^{i+1}\right) = 0 && \text{on } \Gamma_T^i,\\
&\partial_{\nu_i} u_\ell^i = 0 && \text{on } \mathcal{T}_T^i,\\
&u_\ell^i = u_0^i && \text{on } \Omega_0^i,
\end{aligned} \tag{3.19}$$

$i = 1, \dots, k$. Recall again that $\Gamma^0 \equiv \emptyset$, $\psi^k(u) \equiv u$, and $u^{k+1} \equiv 0$.

Assume that the nonnegative functions $u_{\ell-1}^i$ are sufficiently smooth, say,

$$u_{\ell-1}^i \in L^2(0,T; H^1(\Omega^i)) \cap L^\infty(\Omega_T^i)$$

for any $i \in \{1, \dots, k\}$, and let a positive constant c^1 be taken in such a way that

$$0 \le u_{\ell-1}^1 \le c^1$$

a.e. on Ω_T^1, and at the same time

$$0 \le u_{\ell-1}^i \le c^i \equiv \psi^{i-1}(c^{i-1}) \tag{3.20}$$

a.e. on Ω_T^i for each $i \in \{2, \dots, k\}$.

2. According to standard existence theory (Galerkin's method or implicit time discretization method) (see, e.g., [7] for existence and [8] for uniqueness) there exists a unique weak solution u_ℓ^1 satisfying (3.19) and then repeatedly u_ℓ^i for $i = 2, \dots, k$ such that

$$u_\ell^i \in L^2(0,T; H^1(\Omega^i)) \quad \text{and} \quad \partial_t b^i(u_\ell^i) \in L^2(0,T; H^{-1}(\Omega^i)).$$

3. Due to the comparison principle we now show that

$$(3.20) \quad \text{implies} \quad 0 \le u_\ell^i \le c^i$$

for any $i \in \{1, \dots, k\}$. First, observe formally that

$$\begin{aligned}
&\partial_t b^i(u_\ell^i) = \Delta u_\ell^i && \text{in } \Omega_T^i,\\
&\partial_{\nu_i} u_\ell^i + n u_\ell^i = n\psi^{i-1}(u_\ell^{i-1}) \ge 0 && \text{on } \Gamma_T^{i-1},\\
&\partial_{\nu_i} u_\ell^i + n\psi^i(u_\ell^i) = n u_{\ell-1}^{i+1} \ge 0 && \text{on } \Gamma_T^i,\\
&\partial_{\nu_i} u_\ell^i = 0 && \text{on } \mathcal{T}_T^i,\\
&u_\ell^i = u_0^i \ge 0 && \text{on } \Omega_0^i
\end{aligned}$$

and

$$\begin{aligned}
&\partial_t\left(b^i(u_\ell^i) - b^i(c^i)\right) = \Delta\left(u_\ell^i - c^i\right) && \text{in } \Omega_T^i,\\
&\partial_{\nu_i}(u_\ell^i - c^i) + n\left(u_\ell^i - c^i\right) = n\left(\psi^{i-1}(u_\ell^{i-1}) - \psi^{i-1}(c^{i-1})\right) \le 0 && \text{on } \Gamma_T^{i-1},\\
&\partial_{\nu_i}(u_\ell^i - c^i) + n\left(\psi^i(u_\ell^i) - \psi^i(c^i)\right) = n\left(u_{\ell-1}^{i+1} - c^{i+1}\right) \le 0 && \text{on } \Gamma_T^i,\\
&\partial_{\nu_i}(u_\ell^i - c^i) = 0 && \text{on } \mathcal{T}_T^i,\\
&u_\ell^i - c^i \le 0 && \text{on } \Omega_0^i.
\end{aligned}$$

Consequently,

$$0 \le u_\ell^i \le c^i \tag{3.21}$$

a.e. on Ω_T^i for each $i \in \{1, \dots, k\}$. It is not difficult to make it precise like in the proof of Theorem 3.3(ii) above and we omit further details.

4. We propose now to send ℓ to infinity and to show that a subsequence of our solutions u_ℓ of the approximate problems (3.19) converges to a weak solution of (3.6). For this we will need some uniform estimates. We follow the ideas of Alt and Luckhaus [1], but we have to deal with the additional integrals on Γ^i. First we obtain an a priori estimate by testing (3.19) with u_ℓ^i. Denoting

$$\Phi^i(z) \equiv \int_0^z b^i(\xi) d\xi \quad \text{and} \quad B^i(z) \equiv b^i(z) z - \Phi^i(z), \tag{3.22}$$

the parabolic term can be treated with the use of [1, Lemma 1.5]. Performing some calculations we find

$$\begin{aligned} \sum_{i=1}^k \bigg\{ \int_{\Omega^i} B^i(u_\ell^i(x,t))\, dx - \int_{\Omega^i} B^i(u_0^i(x))\, dx + \int_0^t \int_{\Omega^i} \left|\nabla u_\ell^i\right|^2 dx d\tau \\ + n \int_0^t \int_{\Gamma^i} \left(\psi^i(u_\ell^i) u_\ell^i + |u_\ell^{i+1}|^2\right) d\sigma d\tau \bigg\} \\ = \sum_{i=1}^{k-1} n \int_0^t \int_{\Gamma^i} \left(\psi^i(u_\ell^i) u_\ell^{i+1} + u_{\ell-1}^{i+1} u_\ell^i\right) d\sigma d\tau. \end{aligned}$$

Since $u_\ell^i \in L^2(0,T; H^1(\Omega^i))$ the estimate (3.21) holds on Γ_T^i, too. Hence, the right-hand side is uniformly bounded with respect to ℓ and we arrive at

$$\sum_{i=1}^k \left\{ \int_{\Omega^i} B^i(u_\ell^i(x,t))\, dx + \left\|\nabla u_\ell^i\right\|^2_{L^2(\Omega_t^i)} \right\} \le C_n \tag{3.23}$$

for any ℓ and for a.e. $t \in [0,T]$. In order to go to the limit in the nonlinear terms we need strong convergence. To arrive at an estimate concerning the time variable we test our problem with $\varphi^i = \chi_{[t,t+h]}(\tau) w^i$ for $w^i \in H^1(\Omega^i)$, $i = 1, \dots, k$, $w^{k+1} \equiv 0$,

$$\begin{aligned} \sum_{i=1}^k \bigg\{ \int_{\Omega^i} \left(b^i(u_\ell^i(t+h)) - b_\ell^i(u_\ell^i(t))\right) w^i dx + \int_t^{t+h} \int_{\Omega^i} \nabla u_\ell^i \nabla w^i \; dx d\tau \\ + n \int_t^{t+h} \int_{\Gamma^i} \Big((u_\ell^{i+1} - \psi^i(u_\ell^i)) w^{i+1} + (\psi^i(u_\ell^i) - u_{\ell-1}^{i+1}) w^i \Big) d\sigma d\tau \bigg\} \; = \; 0 \end{aligned}$$

and then put $w^i \equiv u_\ell^i(t+h) - u_\ell^i(t)$. Since the items on Γ^i are bounded almost everywhere, after integration over $[0, T-h]$ one gets the estimate

$$\sum_{i=1}^k \int_0^{T-h} \int_{\Omega^i} \left(b^i(u_\ell^i(t+h)) - b^i(u_\ell^i(t))\right) \left(u_\ell^i(t+h) - u_\ell^i(t)\right) \; dx dt \le ch \tag{3.24}$$

with constant c independent of ℓ. Here boundedness (3.23) of the gradient of u_ℓ^i was used. Now, passing to the limit $\ell \to \infty$, the estimates (3.23) and (3.24) together with the weak formulation of problem (3.19) yield the existence of a subsequence of $\{u^i_\ell\}$ such that

$$\begin{aligned} b^i(u^i_\ell) &\to b^i(u^i) && \text{in } L^1(\Omega^i_T),\\ u^i_\ell &\rightharpoonup u^i && \text{in } L^2(0,T;H^1(\Omega^i)),\\ \partial_t b(u^i_\ell) &\rightharpoonup \partial_t b(u^i) && \text{in } L^2(0,T;H^{-1}(\Omega^i)) \end{aligned}$$

for $i = 1,\dots,k$ (cf. [1]). Note, that due to (3.1) and (3.2), u^i_ℓ converges strongly in $L^1(\Omega^i_T)$ and thus, by interpolation with respect to space variables, also strongly in $L^2(\Gamma^i_T)$. Now, it is not difficult to see that u^i, $i = 1,\dots,k$, is a weak solution of (3.6). □

4. Free boundary value problems. In this section we intend to prove the existence of a solution to our original problem (1.1)–(1.5) with discontinuous $b(u)$ by approximating it by problem (3.6). Since we let ε go to zero and n to infinity in (3.6) our main task is to derive a priori estimates that do not depend on ε, n. The simple method of step 4 in the proof of Theorem 3.2 by testing the relation with u^i is not applicable since the bounds depend on n (cf. (3.23)).

To this end we want to test our system (3.6) with

$$\phi^i(u) = \psi^k \circ \dots \circ \psi^i(u), \qquad i = 1,\dots,k,$$

chosen in such a way that the integrals on Γ^i are nonnegative. Since $\partial_t b^i(u^i_\varepsilon)$ only belongs to the dual $L^2(0,T;H^{-1}(\Omega^i))$ we have to formulate an "integration by parts formula" for that case which is proven by Carrillo [4]. Define

$$G^i_\varepsilon(s) = \int_0^s \phi^i((b^i_\varepsilon)^{-1}(r))\,dr; \tag{4.1}$$

we get the following lemma.

LEMMA 4.1. *Let $\phi \in C^{0,1}(\mathbb{R})$ be monotone, let $b_0 = b(u_0) \in L^1(\Omega)$ such that $G(b_0) \in L^1(\Omega)$, let $u \in L^2(0,T;H^1(\Omega))$, and $b(u) \in L^1(Q_T)$ with derivative $\partial_t b(u) \in L^2(0,T;H^{-1}(\Omega))$. Then*

$$G(b(u)) \in L^\infty(0,T;L^1(\Omega))$$

and, for almost every $t \in [0,T]$,

$$\int_\Omega G\big(b(u(x,t))\big)\,dx - \int_\Omega G\big(b_0(x)\big)\,dx = \int_0^t \langle b(u)_t, \phi(u)\rangle\,d\tau.$$

Proof. The lemma is an adaption of [4, Lemma 4] to our problem. □

For the following estimates recall our convention $u^{k+1}_\varepsilon \equiv 0$, $\psi^k(u) \equiv u$. In view of our assumptions due to (1.8) we may estimate G^i_ε below by

$$G^i_\varepsilon(b^i_\varepsilon(u)) \geq \kappa^{k-i} \int_0^{b^i_\varepsilon(u)} (b^i_\varepsilon)^{-1}(r)\,dr = \kappa^{k-i} \int_0^u z\,db^i_\varepsilon(z) = \kappa^{k-i} B^i_\varepsilon(u), \tag{4.2}$$

where $B^i_\varepsilon(u)$ is defined by (3.22) and above by

$$G^i_\varepsilon(b^i_{0\varepsilon}) \leq K^{k-i} \int_0^{b^i_{0\varepsilon}} (b^i_\varepsilon)^{-1}(r)\,dr \leq \frac{K^{k-i}}{\iota} \int_0^{b^i_{0\varepsilon}} r\,dr = \frac{K^{k-i}}{2\iota}\,|b^i_{0\varepsilon}|^2 \tag{4.3}$$

for nonnegative $b^i_{0\varepsilon}$ since $(b^i_\varepsilon)^{-1}(r) \leq \frac{1}{\iota}\,r$ for $r \geq 0$.

Now we test the problem (3.6) with the abovementioned $\phi^i(u^i_\varepsilon)$. According to (1.8), for $i = 1, \dots, k-1$,

$$\int_0^t \int_{\Gamma^i} (u_\varepsilon^{i+1} - \psi^i(u_\varepsilon^i))\,(\phi^{i+1}(u_\varepsilon^{i+1}) - \phi^i(u_\varepsilon^i))\,d\sigma d\tau$$
$$\geq \kappa^{k-1-i} \int_0^t \int_{\Gamma^i} (u_\varepsilon^{i+1} - \psi^i(u_\varepsilon^i))^2\,d\sigma d\tau\,.$$

Consequently,

$$\sum_{i=1}^k \left\{ \int_{\Omega^i} \left(G_\varepsilon^i(b_\varepsilon^i(u_\varepsilon^i(t))) - G_\varepsilon^i(b_\varepsilon^i(u_{0\varepsilon}^i)) \right) dx + \int_0^t \int_{\Omega^i} D\phi^i(u_\varepsilon^i) \left|\nabla u_\varepsilon^i\right|^2 dx d\tau \right\}$$
$$+ n \sum_{i=1}^k \kappa^{(k-1-i)_+} \int_0^t \int_{\Gamma^i} \left(u_\varepsilon^{i+1} - \psi^i(u_\varepsilon^i) \right)^2 d\sigma d\tau \ \leq \ 0.$$

For bounded initial data $u^i_{0\varepsilon}$ the function $G^i_\varepsilon(b^i_\varepsilon(u^i_{0\varepsilon}))$ is bounded in Ω^i uniformly with respect to ε, too. Thus, regarding (4.2), (4.3), and (1.8) again, we have proved the following a priori estimates.

Lemma 4.2. *Suppose assumptions* (3.1)–(3.5) *hold and let* $b^i_{0\varepsilon} = b^i_\varepsilon(u^i_{0\varepsilon})$ *be uniformly bounded in* $L^2(\Omega^i)$ *for* $i = 1, \dots, k$. *Then the solution* u_ε *from Theorem* 3.2 *fulfills the estimates*

$$\int_{\Omega^i} B_\varepsilon^i(u_\varepsilon^i(x,t))\,dx \leq C_1, \tag{4.4}$$

$$\|\nabla u_\varepsilon^i\|_{L^2(\Omega_T^i)} \leq C_2, \tag{4.5}$$

$$n\,\|u_\varepsilon^{i+1} - \psi^i(u_\varepsilon^i)\|^2_{L^2(\Gamma_T^i)} \leq C_3 \tag{4.6}$$

for all $\varepsilon > 0$ *and* $i = 1, \dots, k$.

In order to obtain strong convergence of u^i_ε as $\varepsilon \to 0$ we need an a priori estimate with respect to the time variable (cf. [1, Lemma 1.9]).

Lemma 4.3. *Let the assumptions of Lemma* 4.2 *be fulfilled and* $0 < h < T$. *Then the estimate*

$$\int_0^{T-h} \int_{\Omega^i} \left(b_\varepsilon^i(u_\varepsilon^i(x,t+h)) - b_\varepsilon^i(u_\varepsilon^i(x,t)) \right) \left(u_\varepsilon^i(x,t+h) - u_\varepsilon^i(x,t) \right) dx dt \leq Ch \tag{4.7}$$

holds for all $\varepsilon > 0$ *and* $i = 1, \dots, k$.

Proof. Inserting $\chi_{(t,t+h)} w^i$, $w^i \in V_i$, as a test function into (3.8) we obtain by means of partial integration in time

$$\sum_{i=1}^k \left(\int_{\Omega^i} \left(b_\varepsilon^i(u_\varepsilon^i(t+h)) - b_\varepsilon^i(u_\varepsilon^i(t)) \right) w^i\,dx + \int_t^{t+h} \int_{\Omega^i} \nabla u_\varepsilon^i \nabla w^i\,dx d\tau \right)$$
$$= - \sum_{i=1}^k n \int_t^{t+h} \int_{\Gamma^i} \left(u_\varepsilon^{i+1} - \psi^i(u_\varepsilon^i) \right) \left(w^{i+1} - w^i \right) d\sigma d\tau.$$

Now we choose $w^i = \phi^i(u^i_\varepsilon(t+h)) - \phi^i(u^i_\varepsilon(t))$, $w^{k+1} \equiv 0$, integrate over $t \in [0, T-h]$, and estimate by means of (1.8)

$$\begin{aligned}
&\sum_{i=1}^{k} \kappa^{k-i} \int_0^{T-h} \int_{\Omega^i} \left(b^i_\varepsilon(u^i_\varepsilon(t+h)) - b^i_\varepsilon(u^i_\varepsilon(t))\right)\left(u^i_\varepsilon(t+h) - u^i_\varepsilon(t)\right)\, dxdt \\
&\qquad \le h \sum_{i=1}^{k} K^{k-i} \int_0^{T-h} \int_{\Omega^i} \left|\nabla [u^i_\varepsilon]_h\right| \left(|\nabla u^i_\varepsilon(t+h)| + |\nabla u^i_\varepsilon(t)|\right)\, dxdt \\
&\qquad\quad + nh \sum_{i=1}^{k} K^{(k-i-1)_+} \int_0^{T-h} \int_{\Gamma^i} \left|[u^{i+1}_\varepsilon - \psi^i(u^i_\varepsilon)]_h(x,t)\right| \\
&\qquad\quad \times \Big(|u^{i+1}_\varepsilon(t+h) - \psi^i(u^i_\varepsilon(t+h))| + |u^{i+1}_\varepsilon(t) - \psi^i(u^i_\varepsilon(t))|\Big)\, d\sigma dt,
\end{aligned}$$

where $[u]_h(t) = h^{-1}\int_t^{t+h} u(\tau)\, d\tau$ denotes the Steklov average of $u(t)$. Observe that if X is a normed space, then the Steklov average has the property

$$\|\, [u]_h \,\|_{L^p(0,T;X)} \le \|u\|_{L^p(0,T;X)}.$$

Hence, we finish our calculation using the a priori estimates of Lemma 4.2,

$$\begin{aligned}
&\sum_{i=1}^{k} \kappa^{k-i} \int_0^{T-h} \int_{\Omega^i} \left(b^i_\varepsilon(u^i_\varepsilon(t+h)) - b^i_\varepsilon(u^i_\varepsilon(t))\right)\left(u^i_\varepsilon(t+h) - u^i_\varepsilon(t)\right)\, dxdt \\
&\qquad \le h \sum_{i=1}^{k} K^{k-i} \, \|\nabla u^i_\varepsilon\|^2_{L^2(\Omega^i_T)} + nh \sum_{i=1}^{k} K^{(k-i-1)_+} \, \|u^{i+1}_\varepsilon - \psi^i(u^i_\varepsilon)\|^2_{L^2(\Gamma^i_T)} \\
&\qquad \le Ch,
\end{aligned}$$

which proves (4.7). □

Now we are in the position to formulate the existence result for our original problem (1.1)–(1.5). To this end we give the exact definition of a weak solution of this problem.

DEFINITION 4.4. *A function $u \equiv (u^1, \dots, u^k) \in L^2(0,T;V)$ is a weak solution of problem (1.1)–(1.5) provided there is a function $w \in b(u)$ with $w^i \in L^2(\Omega^i_T)$, $\partial_t w \in L^2(0,T;\tilde{V}^*)$, and initial values b^i_0 in the sense that*

$$\int_0^T \langle \partial_t w, v\rangle\, dt = -\sum_{i=1}^{k} \int_0^T \int_{\Omega^i} (w^i - b^i_0)\, \partial_t v^i \, dxdt \tag{4.8}$$

holds for any $v = (v^1, \dots, v^k) \in L^2(0,T;\tilde{V})$ with $\partial_t v^i \in L^\infty(\Omega^i_T)$, $v^i(T) = 0$, the following identity

$$\int_0^T \langle \partial_t w, \phi\rangle\, dt \; + \sum_{i=1}^{k} \int_0^T \int_{\Omega^i} \nabla u^i \nabla \phi^i \, dxdt = 0 \tag{4.9}$$

is fulfilled for any $\phi \in L^2(0,T;\tilde{V})$, and

$$u^{i+1} = \psi^i(u^i) \qquad \text{a.e. on } \Gamma^i_T. \tag{4.10}$$

Theorem 4.5. *Let $b^i(v)$, $\psi^i(v)$, $i = 1, \dots, k$, be as defined in section* 1 *fulfilling assumptions* (1.6)–(1.8). *Then for given nonnegative $b_0^i \in L^2(\Omega^i)$ there is a nonnegative weak solution u of the system* (1.1)–(1.5) *in the sense of Definition* 4.4.

If, moreover, $0 \le b_0^i \le C^i$ for $i = 1, \dots, k$ where $c^i = (b^i)^{-1}(C^i)$ fulfills $c^{i+1} = \psi^i(c^i)$, then the solution is bounded a.e. on Ω_T by $0 \le u^i \le c^i$.

Proof. We approximate our problem (1.1)–(1.5) by the regularized problem (3.6). As an initial condition we choose $b_\varepsilon^i(u_{0\varepsilon}^i) = b_{0\varepsilon}^i := R_\varepsilon * b_0^i$, where R_ε is the standard mollifier with respect to x. For $b_0^i \ge 0$ the regularization is nonnegative again, thus, $u_{0\varepsilon}^i$ will also be nonnegative for all $i = 1, \dots, k$. Since $(b_\varepsilon^i)^{-1}$ is Lipschitz continuous we have $u_{0\varepsilon}^i \in V_i \cap L^\infty(\Omega^i)$, hence there is a solution u_ε to the approximate problem due to Theorem 3.2.

Our aim is to let $\varepsilon \to 0$ and $n \to \infty$ simultaneously, e.g., choosing $n = \varepsilon^{-1}$. Assuming such connection between ε and n we omit indication of the dependence of u_ε^i on n. Since we have provided all necessary estimates we can use the standard arguments from [1]. First from (4.5) immediately follows the existence of a subsequence (we write u_ε^i again for all subsequences) with

$$u_\varepsilon^i \rightharpoonup u^i \qquad \text{in } L^2(0,T;V_i) \quad \text{as } \varepsilon \to 0 \tag{4.11}$$

($i = 1, \dots, k$). Because of [1, Lemma 4.4], the estimate (4.4) together with Lemma 4.3 yields for a subsequence

$$b_\varepsilon^i(u_\varepsilon^i) \rightharpoonup w^i \in b^i(u^i) \qquad \text{in } L^1(\Omega_T^i) \quad \text{as } \varepsilon \to 0. \tag{4.12}$$

Now, consider relation (3.8) with test functions $\phi \in L^2(0,T;\tilde{V})$. Then the integrals over Γ^i disappear. Since $\tilde{V}$ is a subspace of $V_1 \times \dots \times V_k$ we have $\partial_t b_\varepsilon(u_\varepsilon) := \partial_t b_\varepsilon^1(u_\varepsilon^1) \times \dots \times \partial_t b_\varepsilon^k(u_\varepsilon^k) \in L^2(0,T;\tilde{V}^*)$ and obtain from (3.8) regarding (4.5)

$$\|\partial_t b_\varepsilon(u_\varepsilon)\|_{L^2(0,T;\tilde{V}^*)} \le C.$$

This yields for a subsequence

$$\partial_t b_\varepsilon(u_\varepsilon) \rightharpoonup \partial_t w \qquad \text{in } L^2(0,T;\tilde{V}^*) \quad \text{as } \varepsilon \to 0. \tag{4.13}$$

Moreover, we need strong convergence of a subsequence $\{u_\varepsilon^i\}$ in order to go to the limit in the items defined on the interfaces Γ^i. Recalling construction (2.1) of $b_\varepsilon^i(u)$, thanks to monotonicity of χ_ε the estimate (4.7) for $b_\varepsilon^i(u)$ also holds for its first item,

$$\int_0^{T-h} \int_{\Omega^i} \left(\beta_\varepsilon^i(u_\varepsilon^i(x,t+h)) - \beta_\varepsilon^i(u_\varepsilon^i(x,t))\right) \left(u_\varepsilon^i(x,t+h) - u_\varepsilon^i(x,t)\right) dxdt \le Ch,$$

and then due to (1.7)

$$\iota \, \|u_\varepsilon^i(x,t+h) - u_\varepsilon^i(x,t)\|_{L^2(\Omega_T^i)}^2 \le Ch$$

for all $h > 0$, $\varepsilon > 0$, and $i = 1, \dots, k$. Hence, together with (4.5) Kolmogoroff's compactness theorem yields strong convergence of a subsequence

$$u_\varepsilon^i \to u^i \qquad \text{in } L^2(\Omega_T^i) \quad \text{as } \varepsilon \to 0. \tag{4.14}$$

Strong convergence on the interfaces Γ^i we obtain by application of the interpolation inequality

$$\|u\|_{L^2(\Gamma_T^i)} \le C \, \|u\|_{L^2(0,T;V_i)}^{1-\theta} \, \|u\|_{L^2(\Omega_T^i)}^{\theta}$$

(see [7, Proposition 2]) to the difference $u^i_\varepsilon - u^i$. Regarding (4.5) and (4.14), this inequality yields

$$u^i_\varepsilon \to u^i \qquad \text{in } L^2(\Gamma^i_T) \quad \text{as } \varepsilon \to 0, \tag{4.15}$$

$(i = 1, \dots, k)$. Since we arranged that $n \to \infty$ if $\varepsilon \to 0$, the estimate (4.6) shows that the limit u^i fulfills our jump condition (4.10).

To show that the limit function is a solution, let $\varepsilon \to 0$ in relations (3.7) and (3.8) for test functions as fixed in Definition 4.4 using (4.11)–(4.13), which yields relations (4.8) and (4.9), respectively.

Finally, consider bounded initial values. Then the regularized initial values are also bounded by $b^i_{0\varepsilon} = b^i_\varepsilon(u^i_{0\varepsilon}) \le C^i$. Since $(b^i_\varepsilon)^{-1}(s) \to (b^i)^{-1}(s)$ as $\varepsilon \to 0$ for every fixed $s \ge 0$, we find bounds c^i_ε such that

$$0 \le u^i_{0\varepsilon} \le c^i_\varepsilon, \qquad c^{i+1}_\varepsilon = \psi^i(c^i_\varepsilon), \qquad \text{and} \quad c^i_\varepsilon \to c^i \quad \text{as } \varepsilon \to 0.$$

Hence, the boundedness assertion $u^i_\varepsilon \le c^i_\varepsilon$ a.e. on Ω^i_T from Theorem 3.2 yields the boundedness assertion for u^i. □

All arguments in the preceding proof to obtain the convergence properties (4.11)–(4.15) remain the same if we fix n, choose test functions from $L^2(0,T;V)$, and let ε pass to zero. This proves the existence of a solution to the free boundary problem with conditions on the interfaces Γ^i as formulated in (3.8).

COROLLARY 4.6. *Suppose assumptions of Theorem* 4.5. *Then there is a weak solution to the free boundary problem* (3.6) *with* $b^i(u)$ *given by* (1.6), *i.e., there are functions* $u \in L^2(0,T;V)$ *and* $w^i \in L^2(\Omega^i_T)$ *with* $\partial_t w^i \in L^2(0,T;V^*_i)$ *fulfilling relations* (4.8) *and*

$$\sum_{i=1}^k \left(\int_0^T \langle \partial_t w^i, \phi^i \rangle \, dt \; + \int_0^T \int_{\Omega^i} \nabla u^i \nabla \phi^i \, dx dt \right)$$
$$+ \; n \sum_{i=1}^k \int_0^T \int_{\Gamma^i} \left(u^{i+1} - \psi^i(u^i) \right) \left(\phi^{i+1} - \phi^i \right) \, d\sigma dt = 0$$

for any $\phi \in L^2(0,T;V)$, $u^{k+1} \equiv \phi^{k+1} \equiv 0$.

Remark. For bounded initial values we obtain bounded solutions. DiBenedetto and Vespri prove in [5] that a bounded solution to (1.1) is locally continuous if it can be approximated in the topology of our approximation by a sequence of local smooth solutions to (1.1) for smooth $b^i_\varepsilon(\cdot)$. Hence, if our approximations u^i_ε were smooth on Ω^i_T, then u^i is continuous in Ω^i_T.

5. Nonlinear diffusion. Finally, let $b^i(u)$ and $\psi^i(u)$ be sufficiently smooth functions defined for $u \in \mathbb{R}$, say $C^2(\mathbb{R})$, which for some constants κ and K satisfy

$$0 < \kappa \le (b^i)', \; (\psi^i)' \le K \tag{5.1}$$

for $i \in \{1, \dots, k\}$, $b^i(0) = \psi^i(0) = 0$, and $|(b^i)''|, |(\psi^i)''| \le K$ on $\mathbb{R}$.

The above can be viewed as monotone continuously differentiable regularization of the graph b^i and the function ψ^i from the beginning. See (1.7) and (1.8) above.

The main aim of this section is to show, at least in this regular case, that the nonlinear diffusion multicomponent system with the jump conditions $u^{i+1} = \psi^i(u^i)$ between two components is uniquely solvable. For one component Alt and Luckhaus [1] prove uniqueness for continuous b under an additional regularity assumption which is quite similar to our condition (5.5). In the case without interfaces Γ^i, however, this condition may be omitted (cf. also the remark after the proof of Theorem 5.2). Carrillo [4, Theorem 14 and Corollary 10] proves uniqueness of the solution without condition (5.5) to even more general equations using Kruzhkov's method of doubling variables. In our case with the contact condition this method seems to be very complicated. Therefore, we have derived a comparison result by solving the dual problem again. But contrary to the proof of Theorem 3.3 we have now the jump condition (5.4) on the interfaces, hence the corresponding dual problem has a transmission condition on Γ^i.

Thus, let us now consider the problem

$$\begin{aligned}
&\partial_t b^i(u^i) = \Delta u^i && \text{in } \Omega^i_T\,,\ i=1,\dots,k,\\
&u^i = \psi^{i-1}(u^{i-1}) \quad \text{and} \quad \partial_{\nu_i} u^i + \partial_{\nu_{i-1}} u^{i-1} = 0 && \text{on } \Gamma^{i-1}_T,\ i=2,\dots,k,\\
&\partial_{\nu_i} u^i = 0 && \text{on } \mathcal{T}^i_T\,,\ i=1,\dots,k,\\
&u^k = 0 && \text{on } \Gamma^k_T,\\
&u^i = u^i_0 \geq 0 && \text{on } \Omega^i_0\,,\ i=1,\dots,k,
\end{aligned} \tag{5.2}$$

where initial data u^i_0 are supposed to be bounded and sufficiently smooth.

Definition 5.1. *We say now that a k-tuple $u = (u^1,\dots,u^k) \in L^2(0,T;V)$ is a weak solution of the system* (5.2) *provided that*

(i)

$$\sum_{i=1}^k \left(\int_{\Omega^i} b^i(u^i(x,t))\, \phi^i(x,t)\, dx - \int_0^t \int_{\Omega^i} \left(b^i(u^i)\partial_\tau \phi^i - \nabla u^i \nabla \phi^i \right) \, dx d\tau \right)$$

$$= \sum_{i=1}^k \int_{\Omega^i} b^i(u^i_0)\, \phi^i(x,0)\, dx \tag{5.3}$$

is fulfilled for any $\phi \in L^2(0,T;\tilde V)$ with $\phi_t \in L^2(\Omega_T)$ (recall that $\phi^{i+1} = \phi^i$ on Γ^i is required above) and a.e. $t \in [0,T]$;

(ii)

$$u^{i+1} = \psi^i(u^i) \qquad \text{a.e. on } \Gamma^i_T, \tag{5.4}$$

$i = 1,\dots,k$; and finally,

(iii) *there exists a positive constant C such that*

$$\sum_{i=1}^k \int_0^{T-h} \int_{\Omega^i} \left| u^i(x,t+h) - u^i(x,t) \right| \, dxdt \ \leq\ Ch \tag{5.5}$$

for $h > 0$.

A subsolution (supersolution) is defined by (5.3) *with equality replaced by $\leq$ $(\geq)$ and $\phi^i \geq 0$ on Ω^i_T for each $i \in \{1,\dots,k\}$.*

Theorem 5.2. *Let u be a subsolution and v a supersolution of the problem* (5.2) *with initial data u_0 and v_0, respectively. Then if $u_0 \leq v_0$, i.e., $u^i_0 \leq v^i_0$ a.e. on Ω^i for any $i \in \{1,\dots .k\}$, it follows that*

$$u^i(x,t) \leq v^i(x,t) \tag{5.6}$$

a.e. on Ω^i_T for any $i \in \{1,\dots,k\}$.

Moreover, if u *and* v *are weak solutions of the problem* (5.2) *with initial functions* u_0 *and* v_0*, respectively, then for almost all* $t \in [0,T]$ *the inequality* (3.11) *holds.*

Proof. Let $0 < t \le T$ be fixed. Take the difference of the integral inequalities satisfied by u and v for sufficiently smooth $\phi^i \ge 0$. Then

$$\begin{aligned}
&\sum_{i=1}^{k}\Big\{\int_{\Omega^i}\big(b^i(u^i(x,t)) - b^i(v^i(x,t))\big)\ \phi^i(x,t)\,dx\\
&\qquad -\int_0^t\int_{\Omega^i}\big(\psi^i(x,u^i(x,\tau)) - \psi^i(x,v^i(x,\tau))\big)\big[\alpha^i(x,\tau)\partial_t\phi^i(x,\tau)\\
&\qquad + \mu^i(x,\tau)\Delta\phi^i(x,\tau)\big]\ dxd\tau\Big\}\\
&\qquad + \sum_{i=1}^{k}\int_0^t\int_{\Gamma^i}\big(\psi^i(x,u^i(x,\tau)) - \psi^i(x,v^i(x,\tau))\big)\\
&\qquad \times\big[\mu^i(x,\tau)\partial_{\nu_i}\phi^i + \mu^{i+1}(x,\tau)\partial_{\nu_{i+1}}\phi^{i+1}\big]\ d\sigma d\tau\\
&\quad \le \sum_{i=1}^{k}\int_{\Omega^i}\big(b^i(u_0^i(x)) - b^i(v_0^i(x))\big)\ \phi^i(x,0)\,dx.
\end{aligned} \tag{5.7}$$

Recall that we require $\phi^i = \phi^{i+1}$ on Γ^i. In (5.7),

$$\psi^i(x,u) \equiv \begin{cases} u & \text{if } x \in \Gamma^{i-1},\\ \psi^i(u) & \text{if } x \in \Gamma^i,\end{cases}$$

and for $x \in \Omega^i$, say

$$\psi^i(x,u) \equiv u\,\omega^i(x) + \psi^i(u)\,\varpi^i(x),$$

where nonnegative functions ω^i, $\varpi^i \in C^\infty(\bar\Omega^i)$ are such that

$$\omega^i(x) + \varpi^i(x) \equiv 1$$

for $x \in \bar\Omega^i$, $\omega^i \equiv 1$ on a neighborhood of Γ^{i-1}, and $\omega^i \equiv 0$ on a neighborhood of Γ^i, $\psi^k(x,u) = u$ for $x \in \bar\Omega^k$ and $\psi^1(x,u) = \psi^1(u)$ on $\bar\Omega^1$. Note that due to (5.4)

$$\psi^{i+1}(x,u^{i+1}(x,t)) = \psi^i(x,u^i(x,t)) \tag{5.8}$$

for any $x \in \Gamma^i$ and $t \in (0,T)$ and similarly for v^i. Finally,

$$\alpha^i(x,\tau) \equiv \frac{b^i(u^i(x,\tau)) - b^i(v^i(x,\tau))}{\psi^i(x,u^i(x,\tau)) - \psi^i(x,v^i(x,\tau))}$$

and

$$\mu^i(x,\tau) \equiv \frac{u^i(x,\tau) - v^i(x,\tau)}{\psi^i(x,u^i(x,\tau)) - \psi^i(x,v^i(x,\tau))}\ .$$

It follows from (5.1) that

$$0 < \kappa \le \alpha^i(x,\tau), \mu^i(x,\tau) \le K$$

for any $(x,\tau) \in \Omega^i_T$, $i = 1, \dots, k$. Now, let R_ε be the standard mollifier in (x,τ) variables and set

$$\alpha^i_\varepsilon(x,\tau) \equiv \frac{b^i(u^i_\varepsilon(x,\tau)) - b^i(v^i_\varepsilon(x,\tau))}{\psi^i(x,u^i_\varepsilon(x,\tau)) - \psi^i(x,v^i_\varepsilon(x,\tau))},$$

$$\mu^i_\eta(x,\tau) \equiv \frac{u^i_\eta(x,\tau) - v^i_\eta(x,\tau)}{\psi^i(x,u^i_\eta(x,\tau)) - \psi^i(x,v^i_\eta(x,\tau))},$$

where $u^i_\varepsilon(x,\tau) = (R_\varepsilon * u^i)(x,\tau)$, $u^i_\eta(x,\tau) = (R_\eta * u^i)(x,\tau)$, and analogously for v^i, $0 < \varepsilon, \eta \ll 1$.

Finally, let $\varphi_{\varepsilon\eta} = (\varphi^1_{\varepsilon\eta}, \dots, \varphi^k_{\varepsilon\eta})$, $\varphi_{\varepsilon\eta} = \varphi_{\varepsilon\eta}(x,s)$ be the solution of the regularized dual problem

$$\begin{aligned}
&\alpha^i_\varepsilon(x,t-s)\partial_s\varphi^i = \mu^i_\eta(x,t-s)\Delta\varphi^i && \text{in } \ \Omega^i_t\,,\ i = 1,\dots,k,\\
&\mu^i_\eta(x,t-s)\partial_{\nu_i}\varphi^i + \mu^{i+1}_\eta(x,t-s)\partial_{\nu_{i+1}}\varphi^{i+1} = 0, && \\
&\text{and} \quad \ \varphi^i = \varphi^{i+1} && \text{on } \ \Gamma^i_t\,,\ i = 1,\dots,k-1,\\
&\varphi^k = 0 && \text{on } \ \Gamma^k_t,\\
&\partial_{\nu_i}\varphi^i = 0 && \text{on } \ \mathcal{T}^i_t\,,\\
&\varphi^i = \varphi^i_0,\ \ 0 \le \varphi^i_0 \le 1 && \text{on } \ \Omega^i \times \{s = 0\},
\end{aligned} \tag{5.9}$$

assuming that $\varphi^{i+1}_0 = \varphi^i_0$ on Γ^i. As we were not able to prove the existence of the solution $\varphi^i_{\varepsilon\eta}$ from $W^{2.1}_2(\Omega^i_t)$, we have to work in the class of weak solutions and we rewrite the inequality (5.7) into the form

$$\begin{aligned}
&\textstyle\sum_{i=1}^k \{\int_{\Omega^i} \big(b^i(u^i(t)) - b^i(v^i(t))\big)\ \varphi^i(x,0)\, dx\\
&+\ \textstyle\int_0^t \int_{\Omega^i} \big(\psi^i(x,u^i) - \psi^i(x,v^i)\big)\, \alpha^i(x,\tau)\, \partial_s\varphi^i(x,t-\tau)\, dxd\tau\\
&+\ \textstyle\int_0^t \int_{\Omega^i} \nabla\varphi^i(x,t-\tau)\, \nabla\left[\mu^i(x,\tau)\big(\psi^i(x,u^i) - \psi^i(x,v^i)\big)\right]\, dxd\tau\} \le\ 0
\end{aligned} \tag{5.10}$$

for any φ^i, $\ \varphi^{i+1} = \varphi^i$ on Γ^i_t, where we have put

$$\phi^i(x,\tau) = \varphi^i(x,t-\tau), \qquad i = 1,\dots,k\,.$$

By the weak formulation of the regularized dual problem (5.9) we understand the following identity:

$$\sum_{i=1}^k \int_0^t \int_{\Omega^i} \big(\alpha^i_\varepsilon\ \partial_s\varphi^i\ \xi^i\ +\ \nabla\varphi^i\ \nabla\left(\mu^i_\eta\ \xi^i\right)\big)\ dxd\tau\ =\ 0 \tag{5.11}$$

for any $\xi = (\xi^1, \dots, \xi^k) \in L^2(0,t;\tilde V)$. In (5.11) we expect to have

$$\varphi^i \in H^1(0,t;L^2(\Omega^i)) \cap L^2(0,t;V_i)\ , \quad 0 \le \varphi^i \le 1.$$

Indeed, the following assertion holds.

Lemma 5.3. *Let $0 < \varepsilon, \eta \ll 1$ be given. Then there exists a weak solution $\varphi_{\varepsilon\eta}$ of* (5.9) *such that*

$$\int_0^t \int_{\Omega^i} \left|\partial_s\varphi^i_{\varepsilon\eta}\right|^2\ dxd\tau + \max_{0\le t\le T} \int_{\Omega^i} \left|\nabla\varphi^i_{\varepsilon\eta}\right|^2 dx\ \le\ C_1(\eta), \tag{5.12}$$

where the positive constant C_1 *depends on*

$$\left\|\partial_t \mu^i_\eta\right\|_{L^\infty(\Omega^i_T)} \qquad \textit{and} \qquad \left\|\nabla \mu^i_\eta\right\|_{L^\infty(\Omega^i_T)}$$

and does not depend on ε. *In addition,*

$$\int_0^t \int_{\Omega^i} \left|\nabla \varphi^i_{\varepsilon\eta}\right|^2 \, dx d\tau \;\leq\; C_2(\varepsilon, \eta), \tag{5.13}$$

where the positive constant C_2 *depends on*

$$\left\|\nabla \mu^i_\eta\right\|_{L^2(\Omega^i_T)} \qquad \textit{and} \qquad \left\|\partial_t \alpha^i_\varepsilon\right\|_{L^1(\Omega^i_T)},$$

and

$$0 \leq \varphi^i_{\varepsilon\eta} \leq 1 \qquad \textit{on} \qquad \Omega^i_T \; .$$

Proof. We shall approximate problem (5.9) by the following way:

$$\begin{aligned}
&\alpha^i_\varepsilon(x, t-s)\, \partial_s \varphi^i = \mu^i_\eta(x, t-s)\, \Delta \varphi^i && \text{in } \Omega^i_t, \\
&\mu^i_\eta(x, t-s)\partial_{\nu_i} \varphi^i + n(\varphi^i - \varphi^{i-1}) = 0 && \text{on } \Gamma^{i-1}_t, \\
&\mu^i_\eta(x, t-s)\partial_{\nu_i} \varphi^i + n(\varphi^i - \varphi^{i+1}) = 0 && \text{on } \Gamma^i_t, \\
&\mu^i_\eta(x, t-s)\partial_{\nu_i} \varphi^i = 0 && \text{on } \mathcal{T}^i_t, \\
&\varphi^i = \varphi^i_0 && \text{on } \Omega^i \times \{s = 0\}.
\end{aligned} \tag{5.14}$$

1. First of all, the existence of a weak solution $\varphi_n = \varphi_{\varepsilon\eta n} \in L^2(0, T; V) \cap H^1_2(0, T; V^*)$ of (5.14) such that $0 \leq \varphi^i_n \leq 1$, $i = 1, \ldots, k$, can be proved in the same way as it was done in the proof of Theorem 3.3 above and we omit further details. Now, testing (5.14) with φ^i_n and adding up through $i = 1, \ldots, k$ we get

$$\begin{aligned}
&\textstyle\sum_{i=1}^k \left\{\kappa \int_{\Omega^i} \left|\varphi^i_n(x,t)\right|^2 dx \;+\; 2\kappa \int_0^t \int_{\Omega^i} \left|\nabla \varphi^i_n\right|^2 dx\, ds\right\} \\
&+\; \textstyle\sum_{i=1}^k 2n \int_0^t \int_{\Gamma^i} \left|\varphi^{i+1} - \varphi^i\right|^2 d\sigma ds \;\leq\; \sum_{i=1}^k K \int_{\Omega^i} \left|\varphi^i_0(x)\right|^2 dx \\
&+\; \textstyle\sum_{i=1}^k \left\{\int_0^t \int_{\Omega^i} \left|\partial_s \alpha^i_\varepsilon\right| \; dx ds \;+\; 2 \left\|\nabla \varphi^i_n\right\|_{L^2(\Omega^i_t)} \left\|\nabla \mu^i_\eta\right\|_{L^2(\Omega^i_t)}\right\}
\end{aligned}$$

and (5.13) for $\varphi^i_{\varepsilon\eta n}$ follows easily due to Gronwall's lemma.

2. Now, assume for a moment that φ^i_n are so smooth that we can test (5.14) with $\partial_s \varphi^i_n$ and perform all necessary manipulations to arrive at

$$\begin{aligned}
&\textstyle\sum_{i=1}^k \int_0^t \int_{\Omega^i} \left(\alpha^i_\varepsilon \left|\partial_s \varphi^i_n\right|^2 + \partial_s \left(\mu^i_\eta |\nabla \varphi^i_n|^2\right) - |\nabla \varphi^i_n|^2 \partial_s \mu^i_\eta\right) dx\, ds \\
&+\; \textstyle\sum_{i=1}^k \left\{2 \int_0^t \int_{\Omega^i} \nabla \varphi^i_n \; \partial_s \varphi^i_n \; \nabla \mu^i_\eta \; dx\, ds + n \int_0^t \int_{\Gamma^i} \partial_s (\varphi^{i+1} - \varphi^i)^2 \, d\sigma ds\right\} \;=\; 0.
\end{aligned}$$

Hence, (5.12) for $\varphi^i_{\varepsilon\eta n}$ follows easily. This can be made precise performing analogous manipulations with

$$\partial^h_s \varphi^i_n \equiv \frac{\varphi^i_n(x, s+h) - \varphi^i_n(x, s)}{h}$$

instead of $\partial_s \varphi^i_n$ and let us omit these modifications.

Finally, as all estimates are independent of n, they hold for limit functions $\varphi^i_{\varepsilon\eta}$ also. □

Now, due to (5.8) we can insert

$$\xi^i = \psi^i(\cdot, u^i) - \psi^i(\cdot, v^i)$$

as a test function in (5.11) that together with (5.10) yields

$$\begin{aligned}\sum_{i=1}^{k} \Big\{ \int_{\Omega^i} & \left(b^i(u^i(t)) - b^i(v^i(t))\right) \ \varphi^i_{\varepsilon\eta}(x,0)\, dx \\ &+ \int_0^t \int_{\Omega^i} \big[\left(\psi^i(x,u^i) - \psi^i(x,v^i)\right)\ (\alpha^i - \alpha^i_\varepsilon)\, \partial_s \varphi^i_{\varepsilon\eta} \\ &+ \nabla \varphi^i_{\varepsilon\eta}\, \nabla \left((\mu^i - \mu^i_\eta)(\psi^i(x,u^i) - \psi^i(x,v^i))\right)\big]\ dx d\tau \Big\} \le\ 0. \end{aligned} \tag{5.15}$$

Before letting $\varepsilon \to 0$ let us note that

$$(5.5) \qquad \text{yields} \qquad \int_0^T \int_{\Omega^i} \left|\partial_t u^i_\varepsilon(x,t)\right|\ dxdt\ \le\ C,$$

where u^i_ε is defined above; cf. (5.8)–(5.9). The same holds true for $\partial_t v^i_\varepsilon$. Therefore, due to (5.1) and what follows we have

$$\left\|\partial_t \alpha^i_\varepsilon\right\|_{L^1(\Omega^i_T)}\ \le\ C \int_0^T \int_{\Omega^i} \left(\left|\partial_t u^i_\varepsilon(x,t)\right| + \left|\partial_t v^i_\varepsilon(x,t)\right|\right)\ dx\, dt\ \le\ C$$

and also

$$\left\|\nabla \mu^i_\eta\right\|_{L^2(\Omega^i_T)} \le C,$$

C being independent of ε, η. Hence, (5.13) remains uniformly bounded and we are ready to let $\varepsilon \to 0$ and afterwards $\eta \to 0$ in (5.15) to get

$$\sum_{i=1}^{k} \int_{\Omega^i} \left(b^i(u^i(x,t)) - b^i(v^i(x,t))\right)\ \varphi^i_0(x)\, dx\ \le\ 0.$$

As this holds for any smooth function φ^i_0, $0 \le \varphi^i_0 \le 1$, it also continues to hold for $\varphi^i_0 = \chi(u^i(x,t) - v^i(x,t))$, and (5.6) follows easily. The rest of the proof is the same as in Theorem 3.3. □

Remark. If $\psi^i(u) \equiv u$ for all $i = 1, \dots, k$ we have $\mu^i = \mu^i_\eta \equiv 1$. In that case we do not need condition (5.5) for uniqueness since the estimates (5.12) and (5.13) do not depend on η; moreover, (5.13) is not needed. Then, however, we have no jumps on the interfaces Γ^i, which is the special case of a solution $u \in L^2(0,T;H^1(\Omega))$ on one component.

We show now that the solution of (5.2), which we get as the limit of the sequence of solutions to (3.6) under the assumption (5.1), indeed satisfies (5.5).

Theorem 5.4. *In addition to the conditions of section* 4 *we assume* $u^i_0 \in V_i \cap W^2_1(\Omega^i)$, (5.1), *and the compatibility condition*

$$u^{i+1}_0 = \psi^i(u^i_0) \qquad \text{on } \Gamma^i, \quad i = 1, \dots, k. \tag{5.16}$$

Then the solution of problem (5.2) *obtained by Theorem* 4.5 *is a solution in the sense of Definition* 5.1. *In particular, this solution satisfies condition* (5.5).

Proof. Consider the approximate problem (3.6). Theorem 3.2 yields the existence of a solution u_n. We want to show that $b^i(u_n^i)$ fulfills condition (5.5) with a constant C independent of n. Since b is continuous we obtain strong convergence $b^i(u_n^i) \to b^i(u^i)$ as $n \to \infty$ in $L_1(\Omega_T^i)$ where u is a solution of (5.2) (see the proof of Theorem 4.5). Then (5.1) yields the assertion.

Hence, let us estimate the difference $b^i(u^i(x,t+h)) - b^i(u^i(x,t))$ in $L^1(\Omega^i)$ for a.e. $t \in [0, T-h]$ where u is a solution of (3.6). We follow the idea of the proof of Theorem 3.3 replacing $v(x,t)$ by $u(x,t+h)$. Then all manipulations remain the same if we are able to manage the item

$$b^i(u^i(x,t_0+h)) - b^i(u^i(x,t_0))$$

at some initial time t_0. To this end we fix $h > 0$ and extend problem (2.6) to the interval $t \in [-h, T]$. Define

$$\tilde{u}(x,t) = \begin{cases} u(x,t) & \text{for } t \in [0,T], \\ u_0(x) & \text{for } t \in [-h,0], \end{cases}$$

and note that $\tilde{u}$ is the solution of

$$\begin{aligned} &\partial_t b^i(\tilde{u}^i) = \Delta \tilde{u}^i + F^i && \text{in } \Omega^i \times (-h,T), \\ &\partial_{\nu_i} \tilde{u}^i + n\left(\tilde{u}^i - \psi^{i-1}(\tilde{u}^{i-1})\right) = f^i && \text{on } \Gamma^{i-1} \times (-h,T), \\ &\partial_{\nu_i} \tilde{u}^i + n\left(\psi^i(\tilde{u}^i) - \tilde{u}^{i+1}\right) = f^i && \text{on } \Gamma^{i} \times (-h,T), \\ &\partial_{\nu_i} \tilde{u}^i = f^i && \text{on } \mathcal{T}^{i} \times (-h,T), \\ &\tilde{u}^i = u_0^i && \text{on } \Omega^{i} \times (-h,0], \end{aligned} \tag{5.17}$$

with

$$\begin{aligned} F^i(x,t) &= -\chi_{[-h,0]}(t)\, \Delta u_0^i(x), && t \in [-h,T],\, x \in \Omega^i, \\ f^i(x,t) &= \chi_{[-h,0]}(t)\, \partial_{\nu_i} u_0^i(x), && t \in [-h,T],\, x \in \partial\Omega^i, \end{aligned}$$

$i = 1, \dots, k$. Here we have used compatibility condition (5.16). Note that F^i and f^i do not depend on the approximation parameter n. Now we proceed as in the proof of Theorem 3.3 for problem (5.17) instead of (3.6) with $u = \tilde{u}(x,t), v = \tilde{u}(x,t+h)$. Since we integrate over $\tau \in [-h,t]$, by our construction the item

$$\begin{aligned} \int_{\Omega^i} \left(b^i(u_0^i) - b^i(v_0^i)\right) \varphi^i(x,t)\, dx &:= \int_{\Omega^i} \left(b^i(\tilde{u}^i(x,-h)) - b^i(\tilde{u}^i(x,0))\right) \varphi^i(x,t)\, dx \\ &= \int_{\Omega^i} \left(b^i(u_0^i) - b^i(u_0^i)\right) \varphi^i(x,t)\, dx \end{aligned}$$

on the right-hand side of (3.12) disappears. On the other hand, the following additional items appear:

$$\begin{aligned} r^i(\varphi^i, h) = &\int_{-h}^t \int_{\Omega^i} \left(F^i(x,\tau) - F^i(x,\tau+h)\right)\, \varphi^i(x,t-\tau)\, dx d\tau \\ &+ \int_{-h}^t \int_{\partial\Omega^i} \left(f^i(x,\tau) - f^i(x,\tau+h)\right)\, \varphi^i(x,t-\tau)\, d\sigma d\tau. \end{aligned}$$

Again we choose a solution φ^i of the regularized dual problem (3.16) as test function. Since it is bounded by (3.17) by the definition of F^i, f^i we can estimate these additional items by

$$|r^i(\varphi^i,h)| \le h\,(\|\Delta u_0^i\|_{L^1(\Omega^i)} + \|\partial_{\nu_i} u_0^i\|_{L^1(\partial\Omega^i)}) \equiv C_0^i\,h\,.$$

Therefore, performing the manipulations as in the proof of Theorem 3.3, we finally arrive at

$$\sum_{i=1}^{k} \int_{\Omega^i} |b^i(u^i(x,t+h)) - b^i(u^i(x,t))|\,dx \ \le\ \sum_{i=1}^{k} C_0^i\,h \tag{5.18}$$

for a.e $t \in [0, T-h]$, which concludes the proof. □

Remark. Actually, we have proven a stronger condition than (5.5). Indeed, the solution u_n of the approximate problem (3.6) fulfills the Lipschitz condition (5.18) pointwise with respect to time. Since convergence of $b^i(u_n^i)$ in $L^1(\Omega_T^i)$ implies convergence of a subsequence in $L^1(\Omega^i)$ for a.e. $t \in [0, T-h]$ we even obtain

$$\sum_{i=1}^{k} \int_{\Omega^i} |u^i(x,t+h) - u^i(x,t)|\,dx \ \le\ C\,h \qquad \text{for a.e. } t \in [0, T-h].$$

Acknowledgments. We thank Volkswagen-Stiftung for its support and hospitality through the RiP program at Oberwolfach.

REFERENCES

[1] H. W. Alt and S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311–341.

[2] D. G. Aronson, *The porous medium equation*, in Nonlinear Diffusion Problems (Montecatini Terme, 1985), Lecture Notes in Math. 1224, Springer-Verlag, Berlin, 1986, pp. 1–46.

[3] D. G. Aronson, M. G. Crandall, and L. A. Peletier, *Stabilization of solutions of a degenerate nonlinear diffusion problem*, Nonlinear Anal., 6 (1982), pp. 1001–1022.

[4] J. Carrillo, *Entropy solutions for nonlinear degenerate problems*, Arch. Ration. Mech. Anal., 147 (1999), pp. 269–361.

[5] E. DiBenedetto and V. Vespri, *On the singular equation $\beta(u)_t = \Delta u$*, Arch. Ration. Mech. Anal., 132 (1995), pp. 247–309.

[6] J. Filo, *Finite time of stabilization in the one–dimensional problem of non-steady filtration*, Math. Methods Appl. Sci., 19 (1996), pp. 529–554.

[7] J. Filo and J. Kačur, *Local existence of general nonlinear parabolic systems*, Nonlinear Anal., 24 (1995), pp. 1597–1618.

[8] J. Filo and S. Luckhaus, *Modelling surface runoff and infiltration of rain by an elliptic-parabolic equation coupled with a first order equation on the boundary*, Arch. Ration. Mech. Anal., 146 (1999), pp. 157–182.

[9] J. Kačur, *On a solution of degenerate elliptic-parabolic systems in Orlicz-Sobolev spaces*, Math. Z., 203 (1990), pp. 153–171.

[10] J. Kačur and R. Van Keer, *On a numerical method for a class of parabolic problems in composite media*, Num. Methods Partial Differential Equations, 9 (1993), pp. 711–731.

[11] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Uralceva, *Linear and Quasilinear Equations of Parabolic Type*, Trans. Math. Monogr. 23, AMS, Providence, RI, 1968.

[12] R. Manitz, W. Lucht, K. Strehmel, R. Weiner, and R. Neubert, *On mathematical modeling of dermal and transdermal drug delivery*, J. Pharmaceutical Sciences, 87 (1998), pp. 873–879.

[13] A. M. Meirmanov, *The Stefan Problem*, Walter de Gruyter, New York, 1992.

[14] W. P. Ziemer, *Weakly Differentiable Functions. Sobolev Spaces and Functions of Bounded Variation*, Springer–Verlag, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 33, No. 6, pp. 1455–1482

A RENORMALIZATION METHOD FOR MODULATIONAL STABILITY OF QUASI-STEADY PATTERNS IN DISPERSIVE SYSTEMS[*]

KEITH PROMISLOW[†]

Abstract. We employ global quasi-steady manifolds to rigorously reduce forced, linearly damped dispersive partial differential equations to finite dimensional flows. The manifolds we consider are not invariant, but through a renormalization group method we capture the long-time evolution of the full system as a flow on the manifold. For the parametric nonlinear Schrödinger equation we consider a manifold describing N well-separated pulses and derive an explicit system of ordinary differential equations for the flow on the manifold which captures the leading order pulse motion through the tail-tail interactions. We also outline a rigorous connection between the slow evolution in the hyperbolic PNLS and the fourth-order parabolic phase sensitive amplification equation for fiber optic systems.

Key words. parametric nonlinear Schrödinger, renormalization group, orbital stability, invariant manifold

AMS subject classifications. 34D05, 35P15, 78A60

PII. S0036141000377547

1. Introduction.

Many nonlinear optical processes are modeled by partial differential equations which are dominated by dispersive effects. We are interested in a class of these equations which we write abstractly in terms of a vector field F,

$$U_t = F(U). \tag{1.1}$$

In this setting one can often explicitly construct a family of quasi-steady states $\Phi(\mathbf{p})$ parameterized by $\mathbf{p} \in \mathcal{K} \subset \mathbf{R}^N$ for which the residual vector field $F(\Phi(\mathbf{p}))$ satisfies $F(\Phi(\mathbf{p})) = O(\delta)$ for some $\delta \ll 1$. Such families naturally arise as leading order terms in formal asymptotic expansions of exact steady states or of quasi-steady solutions, as in the example we consider here of a linear sum of steady pulses interacting weakly through asymptotically flat tails. If the underlying steady states possess some stability, it is natural to expect that the manifold $\mathcal{M} = \{\Phi(\mathbf{p}) \big| \mathbf{p} \in \mathcal{K}\}$ will retain some degree of local attractivity, that is, solutions U of the full equation may be decomposed as $U(t) = \Phi(\mathbf{p}(t)) + W(t)$, where the time dependent parameters $\mathbf{p}(t)$ shadow the slow evolution of U along the manifold up to a small remainder term W. Previously, modulational stability applied to individual pulses [35, 24] has yielded results that are local in the sense that the full solution must remain close to the initial pulse configuration for the modulational description to retain its force. The method presented here is not tied to a particular local coordinate system; rather, in the spirit of Goldenfeld and Onoo's renormalization group techniques [13], we build our description by taking an envelope of a family of approximate or "naive" perturbation expansions, renormalizing away secularities through a slow modulation of parameters. In this manner we obtain a rigorous reduction of the infinite dimensional system to

[*] Received by the editors September 1, 2000; accepted for publication (in revised form) December 26, 2001; published electronically July 9, 2002. This work was performed while the author was supported by NSERC grant 611255.

http://www.siam.org/journals/sima/33-6/37754.html

[†] Department of Mathematics, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada (kpromisl@sfu.ca).

a finite dimensional one which is more amenable to analysis. In the application to interacting pulses, the reduced system governs the motion of the pulse positions. The method presented here is robust; without modification it can accommodate the background noise ubiquitous in optical systems, capturing these effects as time dependent perturbations to the reduced family of differential equations.

Invariant manifolds have long played a central role in the reduction of *dissipative* flows to lower dimensional systems. The smoothing properties afforded by dissipation make possible the application of a very developed body of invariant manifold theorems [4, 5, 12] to a wide class of problems, including blow-up, decay, and meta-stable behaviors [6, 10, 23, 32, 33]. For systems which are dominated by dispersion, the construction of invariant manifolds is less clear. Smoothing properties for dispersive systems are typically manifested only in weighted spaces, and results have largely been restricted to single hump traveling waves [34]. The renormalization group technique developed herein overcomes the lack of smoothing properties of the underlying equations by renormalizing away secular terms. In some sense the results we obtain are a step towards a rigorous justification of the Lagrangian reduction or collective coordinate literature widely used to model dispersive Hamiltonian systems; see [3, 18] and particularly [26] and the references therein. These methods assume a multiparameter ansatz and develop a family of ordinary differential equations for the parameters from an averaged variational principle for the Lagrangian (see Whitham [36]), which serves to project the vector field onto the tangent plane of the manifold formed by the parameterized ansatz.

Our technique can also be viewed as a modification of the renormalization group developed by Bricmont and Kupiainen [8] for asymptotics of decay in dissipative partial differential equations. The underpinning of the renormalization group is a *family* of decompositions $\{(G_n, \pi_n, t_n)\}_{n=0}^{\infty}$ where G_n represents a rescaling of the dependent and independent variables, π_n is a projection, and t_n is an initial time. The decompositions serve to break the initial value problem for (1.1) into an *equivalent* sequence of initial value problems on the time intervals $[t_n, t_{n+1}]$. The purpose of renormalization is to adapt the coordinate system to the flow on the given time interval. The projection π_n decomposes the phase space, associating to each point U a point Φ_n in the low dimensional manifold $\Phi_n = \pi_n G_n U$ and a remainder $W_n = (I - \pi_n) G_n U$, so that

$$G_n U = \Phi_n + W_n. \tag{1.2}$$

The decomposition applied to (1.1) yields an evolution equation,

$$\begin{aligned} \partial_t \Phi_n &= \pi_n F_n(\Phi_n + W_n), \\ \Phi_n(0) &= \pi_n G_n U(t_n), \\ \partial_t W_n &= (I - \pi_n) F_n(\Phi_n + W_n), \\ W_n(0) &= (I - \pi_n) G_n U(t_n), \end{aligned} \tag{1.3}$$

where F_n represents the rescaled flow, i.e., $(G_n U)_t = F_n(G_n U)$. The nonlinear semigroup S_n for the nth flow gives the map

$$S_n(t) : \begin{pmatrix} \Phi_n(0) \\ W_n(0) \end{pmatrix} \mapsto \begin{pmatrix} \Phi_n(t) \\ W_n(t) \end{pmatrix},$$

which, together with the rescaling and projection implicit in the mapping

$$U(t_{n+1}) \mapsto \begin{pmatrix} \Phi_{n+1}(0) \\ W_{n+1}(0) \end{pmatrix},$$

induce the renormalization operators

$$R_n : \begin{pmatrix} \Phi_n(0) \\ W_n(0) \end{pmatrix} \mapsto \begin{pmatrix} \Phi_{n+1}(0) \\ W_{n+1}(0) \end{pmatrix}.$$

The renormalization operators enjoy the group structure

$$S(\tau_{n+1}) = G_{n+1}^{-1} \circ R_n \circ R_{n-1} \circ \cdots \circ R_0 \circ G_0,$$

where S denotes the nonlinear semigroup of the original system (1.1). In applications of this method to decay of solutions of dissipative partial differential equations, the rescaling is chosen to follow the self-similar nature of the decay, the remainder is driven to zero, and the renormalization groups R_n converge to a limit R_∞; see [9, 7]. The long-time asymptotics of the original partial differential equation are then reduced to a study of the fixed points of R_∞. In the applications cited, the projection step is performed when studying the fixed points of the renormalization maps, where the projection π_n maps onto a stable manifold of R_n. In the applications we consider here the projection plays a dominant role, and the rescaling is less useful. The renormalization operators R_n do not converge to a limit and we adapt the coordinate system not in response to changes in length and time scales in the underlying evolution, but rather to follow the drift along the quasi-steady manifold, updating the corresponding linearized flow and associated projections which dictate the local evolution.

We assume the quasi-steady manifold is a smoothly parameterized N dimensional manifold $\mathcal{M} = \{\Phi(\mathbf{p}) \big| \mathbf{p} \in \mathcal{K}\}$, and at each point $\Phi(\mathbf{p})$ on the manifold the local linearized operator, $L_{\mathbf{p}}$, engenders a decomposition $X = X_{\mathbf{p}} \oplus Y_{\mathbf{p}}$ of the underlying phase space X into two $L_{\mathbf{p}}$ invariant parts: an N dimensional space $Y_{\mathbf{p}}$ associated with small eigenvalues of $L_{\mathbf{p}}$ and a complimentary space $X_{\mathbf{p}}$ on which $L_{\mathbf{p}}$ generates a $\mathcal{C}_0$ semigroup with a uniform exponential decay rate for all $\mathbf{p}$. This later assumption amounts to a form of normal hyperbolicity, with the fast time scales associated with the decay into a thin neighborhood of the manifold and the slow time scales describing the evolution of the parameters. The flow local to $\Phi(\mathbf{p})$ is governed by the linearized operator, the small residual vector field $F(\Phi)$, and the higher order nonlinearities. In particular the manifold need not be locally invariant under the flow. Rather we assume the manifold is compatible with the local flow in the sense that the space $Y_{\mathbf{p}}$ is well approximated by the local tangent space of the manifold. This decomposition underlies our renormalization group methods and we describe the flow local to $\mathbf{p}$ in terms of a family of modulational equations for the manifold parameters $\mathbf{p}$ and a partial differential equation for the remainder variable W governing the distance to the manifold. We write the evolution equation for W in terms of the linearized operator $L_{\mathbf{p}_0}$ frozen at a point $\mathbf{p}_0$ on the manifold, and build estimates on the growth and decay of W. As the manifold parameters $\mathbf{p}$ evolve away from $\mathbf{p}_0$ a natural secular growth is seen in the estimates for W, and after a finite time, control of W is lost. We remove this secular growth with a renormalization of the evolution equations, updating the base point $\mathbf{p}_0$ through a nonlinear projection; see Figure 2.2 for a graphical representation.

The series of initial value problems generated by the renormalization group method permit us to follow the flow on the manifold without the complication of a time dependent linearized operator. Indeed, we require the operators $L_{\mathbf{p}}$ to generate only a $\mathcal{C}_0$ semigroup on $X_{\mathbf{p}}$, with eventual exponential decay after initial transient growth. It is well known that even if for each fixed t_0 the operator $L(t_0)$ generates an asymptotically contractive semigroup, the flow governed by the linear family of time dependent operators,

$$W_t = L(t)W,$$

need not generate an asymptotically contractive semigroup. The transient growth associated with each fixed operator may never settle down, and the result may be a diminished decay rate or resonant growth. It is precisely this latter phenomenon which Kato excludes with his "stability" assumption (see (1.1) of [17]) required to obtain uniform decay estimates for the semigroup generated by a time dependent family of linear operators. However the verification of the stability assumption typically requires smoothing estimates on the individual operators $L(t_0)$. We use the renormalization method to exploit the fact that the evolution of the parameters $\mathbf{p}$ in $L_{\mathbf{p}}$ is on a slower time scale and the fact that the difference $L_{\mathbf{p}_1} - L_{\mathbf{p}_2}$ is a lower order operator than either of $L_{\mathbf{p}_1}$ or $L_{\mathbf{p}_2}$. In this manner we attain uniform decay estimates without smoothing properties for the semigroup. The situation is particularly clear when considering the evolution of a single pulse, in which case one may be tempted to decompose the solution U of (1.1) as $U(x,t) = \Phi(y) + V(y,t)$, where $y = x - s(t)$ is a traveling variable that shadows the pulse position $s(t)$. When the remainder V is advected with the pulse there is the advantage that the evolution for V is governed by a time independent linear operator

$$V_t = LV + \mathcal{N}(V) + \Phi'(y)s' + V_y s',$$

where $'$ denotes differentiation of a function of one variable, and $\mathcal{N}$ represents nonlinear terms in V. However the term V_y is unbounded and the necessary estimates on V cannot be closed without some smoothing estimates for the semigroup generated by L. It is precisely this "small" infinity $V_y s'$ which we renormalize away.

In section 2 we present the renormalization method for a general framework of dispersive equations, emphasizing the nature of the assumptions required in hypotheses (H0)–(H4). The results of section 2 are summarized in Theorem 2.1. In section 3 we consider applications to the parametrically forced nonlinear Schrödinger (PNLS) equation, which models dispersive phenomenon in damped, forced systems in a variety of settings including plasma waves, Faraday resonance, spin waves and magnetic solitons in ferro-magnets, and pattern formation in optical parametric oscillators [2, 25]. By considering a model for which the linearized operators have been studied [25], we considerably simplify the presentation of the results. Specifically we show the stability of fronts to time dependent perturbations in the defocusing case, and describe the evolution of trains of N well-separated pulses in the focusing case, including pulses with differing up-down orientations. For each configuration of pulses we obtain a family of differential equations for the pulse positions which show that like-signed pulses attract, while opposite-signed pulses repel. This result is evocative of the meta-stable pattern evolution obtained by Carr and Pego [11] for fronts in a reaction-diffusion equation. We contrast our results with those of Kapitula and Sandstede [19], who study the stability of *exact* N-pulses in PNLS created in an orbit-flip bifurcation under the addition of a dissipative regularizing term. We study the unregularized problem, and the modulational equations for the pulse positions (3.53) show there are no exact N-pulse solutions with wide pulse spacing for PNLS. In the discussion we sketch an argument which shows that to leading order the meta-stable pulse motion in the focusing PNLS is the same as that in the phase sensitive amplification (PSA) equation for fiber optics; see [22, 30]. We mention that an extension of the work on N-pulses [27] which recovers some of the results presented here is in preparation [28].

We will employ the following notation throughout the paper. The usual L^p norm is denoted $\|\cdot\|_p$, and the L^2 inner product $(\cdot,\cdot)_2$. The Sobolev norms in H^s will be denoted $\|\cdot\|_{H^s}$, while the induced operator norms on L^2 and H^2 are denoted $\|\cdot\|_{*,2}$ and $\|\cdot\|_{*,H^s}$, respectively. A superscript * denotes complex conjugation while $\Re z$ and $\Im z$ denote the real and imaginary parts of z. The adjoint of an operator L with respect to the L^2 inner product is denoted $L^\dagger$. The superscript ^{t} denotes transposition, and $^\perp$ denotes the orthogonal complement in L^2. The spectrum of a linear operator L is denoted $\sigma(L)$, which we divide into essential $\sigma_e(L)$ and point $\sigma_p(L)$ spectrum. The resolvent set is denoted by $\varrho(L)$. The range and kernel of L are denoted $\mathcal{R}(L)$ and $\ker(L)$, respectively. A superscript $'$ will denote differentiation of a function of a single variable with respect to that variable. We will denote by M any positive constant whose value, which may change from line to line, is independent of any of the small parameters. Occasionally for clarity, we will carry the same constant from one line to the next; we will denote such constants by $M_1, M_2, \ldots$.

2. Abstract formulation of the renormalization method. We consider a class of damped, dispersive wave equations

$$U_t = F(U), \tag{2.1}$$

where

$$F(U) = JU_{xx} + f(U) \tag{2.2}$$

for J a skew matrix. The dependent variable U takes values in $\mathbf{R}^d$, $U : \mathbf{R}\times\mathbf{R}_+ \mapsto \mathbf{R}^d$, and the nonlinear term $f \in \mathcal{C}^2(\mathbf{R}^d, \mathbf{R}^d)$. Of fundamental interest is the evolution of the solutions which reside in a neighborhood of a manifold $\mathcal{M}$ smoothly parameterized by steady or quasi-steady solutions $\Phi(x, \mathbf{p})$ of (2.1) for $\mathbf{p} \in \mathcal{K}$, a compact subset of $\mathbf{R}^N$. More specifically we require that the residual vector field $F\big|_{\mathcal{M}}$ satisfies

$$F(\Phi(\mathbf{p})) = 0 \qquad \text{or} \qquad \|F(\Phi(\mathbf{p}))\|_{H^1} = O(\delta(\mathbf{p})), \tag{2.3}$$

where $\delta(\mathbf{p}) \le \delta_0 \ll 1$ for $\mathbf{p} = (p_1, \ldots, p_N)^t \in \mathcal{K}$. To increase the scope of applications of our method, we modify the equations under consideration to include a small forcing term

$$U_t = F(U) + \delta_0 \tilde{\xi}(U, x, t), \tag{2.4}$$

where $\tilde{\xi}$ satisfies the bound

$$\|\tilde{\xi}(U(\cdot,t),\cdot,t)\|_{H^1} \le M\left(1 + \|U\|_{H^1}^p\right) \tag{2.5}$$

for some $M > 0$ and positive integer p.

The dynamics of (2.4) local to the manifold $\mathcal{M}$ are dominated by the linearized flow. Writing the solution U as a sum

$$U(x,t) = \Phi(x, \mathbf{p}(t)) + W(x,t) \tag{2.6}$$

the evolution for the remainder W is given by

$$W_t = L_{\mathbf{p}} W + \mathcal{N}(W) - \nabla_{\mathbf{p}} \Phi(\mathbf{p})\mathbf{p}' + \delta_0 \xi, \tag{2.7}$$

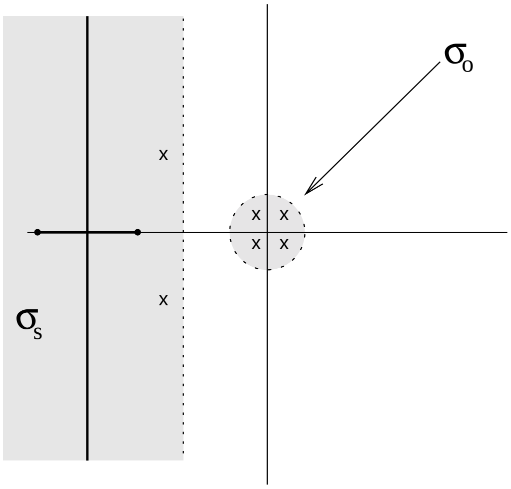


FIG. 2.1. *The spectral decomposition.* $\times$*'s denote eigenvalues and solid dots denote branch points. The dotted line represents the boundary* $\Re\lambda = -k$ *of* σ_s *and the dotted circle of radius* δ_0 *encloses the small eigenvalues of* σ_0.

where

$$L_{\mathbf{p}} = J\partial_x^2 + \nabla f(\Phi(\mathbf{p})), \tag{2.8}$$

the nonlinear terms, $\mathcal{N}$, take the form

$$\mathcal{N}(W) = f(\Phi(\mathbf{p}) + W) - f(\Phi(\mathbf{p})) - \nabla f(\Phi(\mathbf{p}))W,$$

and $\xi \equiv \tilde{\xi} + F(\Phi)/\delta_0$ also satisfies (2.5). The family of linearized operators $\{L_{\mathbf{p}}\}_{\mathbf{p}\in\mathcal{K}}$ plays a central role in the analysis which follows. We make the following assumptions (H0)–(H4) about the operators and the quasi-stationary manifold $\mathcal{M}$.

Quasi-stationarity.

(H0) The manifold $\mathcal{M} = \{\Phi(\mathbf{p}) \big| \mathbf{p} \in \mathcal{K}\}$ is quasi-steady in the sense that

$$\|F(\Phi(\mathbf{p}))\|_{H^1} \leq M\delta_0 \tag{2.9}$$

for some $M > 0$ and all $\mathbf{p} \in \mathcal{K}$. Moreover the quasi-steady ansatz Φ and its first two derivatives with respect to x and $\mathbf{p}$ are uniformly bounded, and the forcing term f and its first two derivatives are uniformly bounded in a neighborhood of $\mathcal{M}$ by some positive constant M.

Normal hyperbolicity.

(H1) The spectrum of each operator $L_{\mathbf{p}}$ may be decomposed into a stable part σ_s, strictly contained in the left-half complex plane, and a slow part σ_0, comprised of a fixed, finite number of small eigenvalues; see Figure 2.1. Specifically

$$\sigma(L_{\mathbf{p}}) = \sigma_s \cup \sigma_0, \tag{2.10}$$

where $\sigma_s \subset \{\lambda \big| \Re\lambda \leq -k\}$ for some $k > 0$ and $\sigma_0 \subset \{\lambda \big| \quad |\lambda| \leq \delta_0\}$, consists of N eigenvalues, up to multiplicity. Both N and k may be chosen independent of $\mathbf{p} \in \mathcal{K}$.

(H2) Each fixed operator $L_{\mathbf{p}}$ generates a $\mathcal{C}_0$ semigroup $S_{\mathbf{p}}$ which satisfies

$$\|S_{\mathbf{p}}(t)u\|_{H^1} \leq Me^{-kt}\|u\|_{H^1} \quad \text{for all } t \geq 0, u \in X_{\mathbf{p}}, \tag{2.11}$$

where $X_{\mathbf{p}}$ is the $L_{\mathbf{p}}$ invariant subspace of H^1 of codimension N associated with the spectrum σ_s. Moreover, M may be chosen independent of $\mathbf{p} \in \mathcal{K}$.

Compatibility. We denote by $Y_{\mathbf{p}}$ the $L_{\mathbf{p}}$ invariant subspace of dimension N complimentary to $X_{\mathbf{p}}$, and refer to $Y_{\mathbf{p}}$ as the slow space.

(H3) We assume that the slow space $Y_{\mathbf{p}}$ is well approximated by the tangent plane of the manifold $\mathcal{M}$; i.e., there is a constant $\delta_C > 0$ small enough and a given ordering $\{\Psi_1, \dots, \Psi_N\}$ of the eigenfunctions of $Y_{\mathbf{p}}$ to which there corresponds a parameterization of the manifold $\mathcal{M}$ verifying

$$\left\| \Psi_i(\mathbf{p}) - \frac{\partial \Phi(\cdot, \mathbf{p})}{\partial \mathbf{p}_i} \right\|_{H^1} \leq \delta_C \qquad \text{for} \quad i = 1, \dots, N \tag{2.12}$$

for each point $\mathbf{p} \in \mathcal{K}$. Here $\delta_C > 0$ is independent of $\mathbf{p} \in \mathcal{K}$.

We denote by $\pi_{\mathbf{p}}$ the $L_{\mathbf{p}}$ spectral projection whose range is $Y_{\mathbf{p}}$, which we construct explicitly as

$$\pi_{\mathbf{p}} u = \sum_{i=1}^{N} (u, \Psi_i^\dagger)_2 \Psi_i, \tag{2.13}$$

where the adjoint eigenvectors $\Psi_i^\dagger$ have been chosen to satisfy the orthonormality conditions

$$(\Psi_i, \Psi_j^\dagger)_2 = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{2.14}$$

We note that the projection π takes this form even if the eigenvalues in σ_0 are not simple.

Stability.

(H4) We assume that the adjoint eigenvectors normalized by (2.14) are uniformly bounded and depend smoothly upon $\mathbf{p}$ for $\mathbf{p} \in \mathcal{K}$; that is,

$$\max_{\substack{i=1,\dots,N \\ \mathbf{p} \in \mathcal{K}}} \left(\|\Psi_i^\dagger(\mathbf{p})\|_{H^1} + \| \quad |\nabla_{\mathbf{p}}^2 \Psi_i^\dagger(\mathbf{p})| \quad \|_{H^1} \right) \leq M. \tag{2.15}$$

We remark that with the normalization we have taken, (2.15) is equivalent to the Jordon chain structure of the eigenvectors of $L_{\mathbf{p}}$ in σ_0 being independent of $\mathbf{p} \in \mathcal{K}$.

Under these assumptions we may prove our main result, stated below.

THEOREM 2.1. *Let a quasi-steady manifold $\mathcal{M} = \{\Phi(\mathbf{p}) \big| \mathbf{p} \in \mathcal{K}\}$ with parameters from $\mathcal{K} \subset \mathbf{R}^N$ be given which satisfies the hypotheses* (H0)–(H4) *for some positive constants M and k. Then for ϵ, δ_0, and δ_C small enough, in terms of M and k, there exist a finite constant M_0 and times T, and T^i all positive, such that for all initial data $U(x, t_0) = U_0(x)$ of the form* (2.20), *the solution U of* (2.4) *can be decomposed as*

$$U = \Phi(x, \mathbf{p}(t)) + W(x, t) \qquad \textit{for}\, t \in [t_0, t_0 + T], \tag{2.16}$$

where the remainder W satisfies

$$\|W(\cdot, t)\|_{H^1} \leq M_0 \left(\epsilon e^{-k(t-t_0)} + \delta_0 \right). \tag{2.17}$$

The parameters $\mathbf{p}(t) = (p_1, \dots, p_N)^t$ *may be chosen to lie on a smooth curve in* $\mathcal{K}$*, and* T*, the time for the curve to exit* $\mathcal{K}$*, satisfies*

$$T > M_0 \operatorname{dist}(\mathbf{p}_*, \partial\mathcal{K})/\delta_0, \tag{2.18}$$

where $\mathbf{p}_*$ *is as given in* (2.20). *After an initial transient, that is, for* $t \geq t_0 + T^i$*, the evolution of the parameters is governed to leading order by the closed system*

$$p_i' = \left(\delta_0 \tilde{\xi} + F(\Phi(\mathbf{p})), \Psi_i^\dagger(\mathbf{p})\right)_2 + O(\delta_0^2) \qquad \text{for } t \geq t_0 + T^i, \tag{2.19}$$

for $i = 1, \dots, N$.

Remark 1. If the evolution equation (2.19) precludes the parameter $\mathbf{p}(t)$ from reaching the boundary $\partial\mathcal{K}$, then we may take $T = \infty$ in the theorem above. We may take $\mathcal{K} \subset \mathbf{R}^N$ to be unbounded so long as the hypotheses (H0)–(H4) hold uniformly for all $\mathbf{p} \in \mathcal{K}$, particularly if M and k in (H0), (H2), and (H4) can be chosen independent of $\mathbf{p}$ over the unbounded set $\mathcal{K}$.

Remark 2. If the initial data U_0 are taken to lie on the manifold, that is, if $U_0 = \Phi(\mathbf{p}_*)$ for some $\mathbf{p}_* \in \mathcal{K}$, then there is no initial transient and we may take $\epsilon = T^i = 0$, and we recover the collective coordinate equations (2.19) immediately at $t = t_0$.

Remark 3. The small parameter δ_C of hypothesis (H3) need only be small enough that the equations (2.28) for $\mathbf{p}'$ have nonzero denominators and that Proposition 2.2 holds. Further reduction in the size of δ_C does not strengthen the results. The error term for the evolution of the manifold parameters (2.19) after the initial transient is dominated by the influence of the remainder. To obtain a more accurate resolution of this reduced flow we must reduce the size of the remainder, which is equivalent to reducing δ_0.

We follow the renormalization group approach outlined in the introduction, breaking the time domain up into intervals $\{[t_n, t_{n+1}]\}_{n=0}^\infty$ and decomposing the space domain with projections $\{\pi_n\}_{n=0}^\infty$. There are two distinct regimes: an initial transient in which the solution converges into a thin $O(\delta_0)$ neighborhood of the manifold and an asymptotic state in which the solution remains within this thin neighborhood; see Figure 2.2. In the asymptotic state the evolution of the full solution can be well described by a closed system of differential equations for the manifold parameters.

2.1. Evolution equations.

We now consider the evolution of a given initial datum U_0 of (2.4), which is near to the manifold $\mathcal{M}$ in the sense that

$$U_0(x) = \Phi(x, \mathbf{p}_*) + \epsilon \hat{W}_0, \tag{2.20}$$

where $\|\hat{W}_0\|_{H^1} \leq 1$ and $\mathbf{p}_*$ is some element of $\mathcal{K}$. It is natural to take ϵ intermediate in size between 1 and δ_0, $1 \gg \epsilon \gg \delta_0$, and we investigate the relaxation of an initial perturbation into the relatively smaller $O(\delta_0)$ neighborhood of the manifold $\mathcal{M}$.

Our analysis requires a local coordinate system in which $W \in X_{\mathbf{p}_0}$ for some $\mathbf{p}_0$. This amounts to solving a nonlinear equation.

Proposition 2.2. *Let the quasi-steady manifold* $\mathcal{M}$ *satisfying hypotheses* (H0)–(H2), (H4) *for some constants* M *and* k *be given. Then for* δ_C *in* (H3) *small enough, in terms of* M *and* k *but independently of* δ_0*, there exist* M_0 *and* $\epsilon_0 > 0$ *such that for all* $\mathbf{p}_* \in \mathcal{K}^o \equiv \{\mathbf{p} \in \mathcal{K} | d(\mathbf{p}_*, \partial\mathcal{K}) > \epsilon_0^{1/2}\}$*, and for all* $\hat{W}_0$ *satisfying* $\|\hat{W}_0\|_{H^1} \leq 1$ *there is a unique smooth function* $\mathcal{H} = \mathcal{H}(\epsilon; \hat{W}_0)$ *which maps* $\mathcal{H} : [-\epsilon_0, \epsilon_0] \to \mathbf{R}^N$ *such that* $\mathcal{H}(0) = 0$ *and the point* $\mathbf{p}(\epsilon) = \mathbf{p}_* + \mathcal{H}(\epsilon)$ *satisfies* $\mathbf{p} \in \mathcal{K}$ *and*

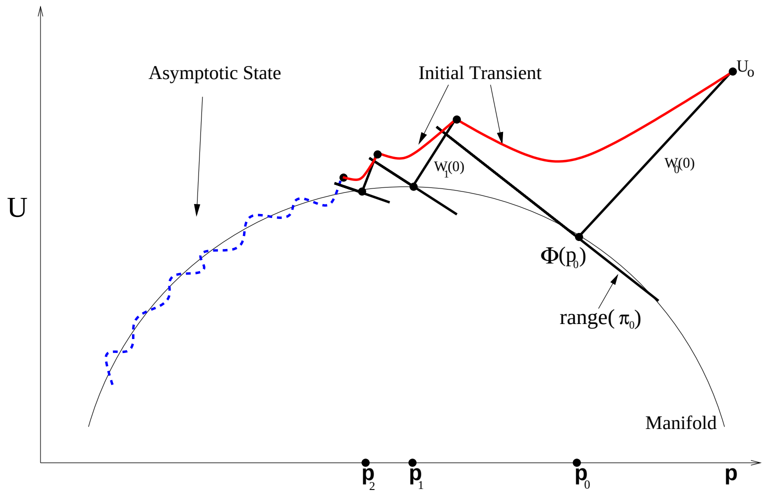


FIG. 2.2. *A sequence of orbits produced by the iterative scheme. The base points* $\mathbf{p}_i$ *are indicated on the horizontal axis. The range of* $\pi_{\mathbf{p}_i}$ *is not the tangent space of the manifold at* $\mathbf{p} = \mathbf{p}_i$ *since* Φ *is not an exact solution of* (2.1). *The orbit initially converges towards the manifold (initial transient—solid) and then remains within an* $O(\delta_0)$ *neighborhood (asymptotic state—dotted).*

$$W_0 \equiv \epsilon \hat{W}_0 + \Phi(\mathbf{p}_*) - \Phi(\mathbf{p}(\epsilon)) \in X_{\mathbf{p}(\epsilon)}. \tag{2.21}$$

Moreover, if $\hat{W}_0 \in X_{\tilde{\mathbf{p}}}$*, then*

$$|\mathbf{p}(\epsilon) - \mathbf{p}_*| \leq M_0 \epsilon |\mathbf{p}_* - \tilde{\mathbf{p}}|. \tag{2.22}$$

Proof. The condition (2.21) is equivalent to

$$0 = \pi_{\mathbf{p}} W_0 = \pi_{\mathbf{p}}(\epsilon \hat{W}_0 + \Phi(\mathbf{p}_*) - \Phi(\mathbf{p})), \tag{2.23}$$

which from (2.13) and the linear independence of the eigenvectors Ψ_i is equivalent to the equations

$$\Gamma_i(\epsilon, \mathbf{p}) \equiv (\epsilon \hat{W}_0 + \Phi(\mathbf{p}_*) - \Phi(\mathbf{p}), \Psi_i^\dagger(\mathbf{p}))_2 = 0 \tag{2.24}$$

for $i = 1, \dots, N$. For $\epsilon = 0$, one solution is $\mathbf{p} = \mathbf{p}_*$. The stability hypothesis (H4) dictates that the adjoint eigenfunctions $\Psi_i^\dagger(\mathbf{p})$ depend smoothly upon the parameters $\mathbf{p}$. We apply the implicit function theorem to find a curve of solutions parameterized by ϵ. We introduce $\Gamma = (\Gamma_1, \dots, \Gamma_N)^t$ and find from (2.12) that the ij entry of the gradient $\nabla_{\mathbf{p}} \Gamma$ satisfies

$$\left(\nabla_{\mathbf{p}} \Gamma \big|_{(0,\mathbf{p}_*)} \right)_{ij} = - \left(\frac{\partial \Phi}{\partial \mathbf{p}_i}, \Psi_j^\dagger \right)_2 = -(\Psi_i, \Psi_j^\dagger)_2 + O(\delta_C). \tag{2.25}$$

From the orthonormality relation (2.14) we may rewrite this as

$$\nabla_{\mathbf{p}} \Gamma \big|_{(0,\mathbf{p}_*)} = -I + O(\delta_C). \tag{2.26}$$

Moreover it is easy to see from (2.24) that

$$\frac{\partial \Gamma}{\partial \epsilon} \bigg|_{(\epsilon,\mathbf{p})} = \begin{pmatrix} (\hat{W}_0, \Psi_1^\dagger(\mathbf{p}))_2 \\ \vdots \\ (\hat{W}_0, \Psi_N^\dagger(\mathbf{p}))_2 \end{pmatrix},$$

while all higher order derivatives of Γ with respect to ϵ are zero. From (2.15) we find that $|\frac{\partial\Gamma}{\partial\epsilon}|_{(0,\mathbf{p})}|$ is bounded uniformly for all $\hat{W}$ in the unit ball of H^1 and all $\mathbf{p} \in \mathcal{K}$ from hypothesis (H4). From (2.26) it follows that $\nabla_{\mathbf{p}}\Gamma\big|_{(0,\mathbf{p}_*)}$ is invertible for δ_C small enough, with the modulus of invertibility $1/|[\nabla_{\mathbf{p}}\Gamma\big|_{(0,\mathbf{p}_*)}]^{-1}|$ uniformly bounded independent of $\hat{W}$ in the unit ball of H^1 and $\mathbf{p} \in \mathcal{K}^o$. Moreover, the second derivatives of Γ with respect to $\mathbf{p}$ are uniformly bounded by hypotheses (H0) and (H4). The implicit function theorem guarantees the existence of a smooth function $\mathcal{H}$ which provides the solution of (2.21). The interval of existence of $\mathcal{H}$, denoted $[-\epsilon_0, \epsilon_0]$, can be chosen independent of $\hat{W}$ and $\mathbf{p}$ due to the uniformity of the bounds on the invertibility of $\nabla_{\mathbf{p}}\Gamma\big|_{(0,\mathbf{p}_*)}$ and the second derivatives of Γ. Moreover, if $\hat{W}_0 \in X_{\tilde{\mathbf{p}}}$, then we have $(\hat{W}_0, \Psi_i^\dagger(\tilde{\mathbf{p}}))_2 = 0$ for $i = 1, \ldots, N$. This yields the estimate

$$|(\hat{W}_0, \Psi_i^\dagger(\mathbf{p}_*))_2| \leq |(\hat{W}_0, \Psi_i^\dagger(\tilde{\mathbf{p}}) - \Psi_i^\dagger(\mathbf{p}_*))_2| = O(|\mathbf{p}_* - \tilde{\mathbf{p}}|).$$

In this case we have the bound $|\frac{\partial\Gamma}{\partial\epsilon}|_{(0,\mathbf{p}_*)}| = O(|\mathbf{p}_* - \tilde{\mathbf{p}}|)$, from which the implicit function theorem yields (2.22). Finally, from the bound (2.22) we see that $\mathbf{p}_* \in \mathcal{K}^o$ implies the curve $\mathbf{p}(\epsilon)$ lies within $\mathcal{K}$ for $|\epsilon| \leq \epsilon_0$ and ϵ_0 small enough. ☐

We employ the proposition above to fix the base point $\mathbf{p}_0 = \mathbf{p}_* + \mathcal{H}(\epsilon)$ and define the initial data W_0 from (2.21). We choose an evolution equation for W which leaves $X_{\mathbf{p}_0}$ positively invariant. With $L_{\mathbf{p}_0}$ as the principle linear operator we rewrite (2.7) as

$$W_t = L_{\mathbf{p}_0}W + \mathcal{N}(W) - K_{\mathbf{p}_0}\mathbf{p}' + B(\mathbf{p}_0, \mathbf{p})W - E(\mathbf{p}_0, \mathbf{p})\mathbf{p}' + \delta_0\xi, \tag{2.27}$$

where $B(\mathbf{p}_0, \mathbf{p}) = L_{\mathbf{p}} - L_{\mathbf{p}_0}$, $K_{\mathbf{p}_0} = (\Psi_1(\mathbf{p}_0), \ldots, \Psi_N(\mathbf{p}_0))$ is a $d \times N$ matrix whose columns span $Y_{\mathbf{p}_0}$, and $E(\mathbf{p}_0, \mathbf{p}) = \nabla_{\mathbf{p}}\Phi(\mathbf{p}) - K_{\mathbf{p}_0}$ controls the difference between the tangent plane of $\Phi(\mathbf{p})$ and $Y_{\mathbf{p}_0}$, which by the compatibility hypothesis (H3) is $O(\delta_C)$. It is important for our analysis that $B(\mathbf{p}, \mathbf{p}_0) = L(\mathbf{p}) - L(\mathbf{p}_0) = \nabla f(\Phi(\mathbf{p})) - \nabla f(\Phi(\mathbf{p}_0))$ is a bounded operator. We note that $\pi_{\mathbf{p}}u = K_{\mathbf{p}}\nu$, where $\nu \in \mathbf{R}^N$ has ith component $\nu_i = (u, \Psi_i^\dagger(\mathbf{p}))_2$.

We now project (2.27) onto the space $X_{\mathbf{p}_0}$, demanding that $\pi_{\mathbf{p}_0}W_t = 0$. This condition yields N equations which determine the N unknowns $\mathbf{p}'$,

$$p_i' = \frac{(\mathcal{N}(W) + B(\mathbf{p}_0, \mathbf{p})W + \delta_0\xi, \Psi_i^\dagger(\mathbf{p}_0))_2}{1 + (E_i(\mathbf{p}_0, \mathbf{p}), \Psi_i^\dagger(\mathbf{p}_0))_2}. \tag{2.28}$$

We supplement these equations with the initial condition

$$\mathbf{p}(t_0) = \mathbf{p}_0. \tag{2.29}$$

The evolution for W may be recast as

$$\begin{aligned} W_t &= L_{\mathbf{p}_0}W + \mathcal{G}, \\ W(t_0) &= W_0, \end{aligned} \tag{2.30}$$

where

$$\mathcal{G} = (I - \pi_{\mathbf{p}_0})\big(B(\mathbf{p}_0, \mathbf{p})W + \mathcal{N}(W) + E(\mathbf{p}_0, \mathbf{p})\mathbf{p}' + \delta\xi\big), \tag{2.31}$$

and W_0 is given by (2.21).

The equation (2.30) admits the mild solution

$$W(t) = S_0(t-t_0)W_0 + \int_{t_0}^{t} S_0(t-t')\mathcal{G}(t')dt', \tag{2.32}$$

where S_0 is the semigroup generated by $L_{\mathbf{p}_0}$. To control the decay of $\|W\|_{H^1}$ and the motion along the manifold $\mathcal{M}$, we introduce the quantities

$$\begin{aligned} T_0(t) &= \sup_{t_0<t'<t} e^{k(t'-t_0)}\|W(t')\|_{H^1}, \\ T_1(t) &= \sup_{t_0<t'<t} |\mathbf{p}(t') - \mathbf{p}_0|. \end{aligned} \tag{2.33}$$

We note that T_0 affords the estimate

$$\|W(t')\|_{H^1} \le e^{-k(t'-t_0)}T_0(t) \qquad \text{for all } t_0 < t' < t. \tag{2.34}$$

The decay estimates of (2.11) applied to the mild solution yield the bound

$$\|W(t)\|_{H^1} \le Me^{-k(t-t_0)}\|W_0\|_{H^1} + M\int_{t_0}^{t} e^{-k(t-t')}\|\mathcal{G}(t')\|_{H^1}dt'. \tag{2.35}$$

From (2.31) it follows that

$$\|\mathcal{G}\|_{H^1} \le \|B(\mathbf{p}_0,\mathbf{p})\|_{*,H^1}\|W\|_{H^1} + \|\mathcal{N}(W)\|_{H^1} + \|E(\mathbf{p}_0,\mathbf{p})\|_{H^1}|\mathbf{p}'| + \delta_0\|\xi\|_{H^1}. \tag{2.36}$$

Since the nonlinearity f is smooth, as is the ansatz function Φ, it follows easily that

$$\|B(\mathbf{p}_0,\mathbf{p})\|_{*,H^1} \le M|\mathbf{p}_0 - \mathbf{p}| \le MT_1. \tag{2.37}$$

Note that under the evolution (2.28) T_1 will grow with time; this is the secularity present in our system which we renormalize away. For $\|W\|_{H^1}$ small the nonlinearity $\mathcal{N}(W)$ and forcing ξ satisfy

$$\|\mathcal{N}(W)\|_{H^1} \le M\|W\|_{H^1}^2, \tag{2.38}$$

$$\|\xi\|_{H^1} \le M\left(1 + \|W\|_{H^1} + \|W\|_{H^1}^2\right). \tag{2.39}$$

From the compatibility assumption we have

$$\begin{aligned} \|E(\mathbf{p}_0,\mathbf{p})\|_{H^1} &\le \|K_{\mathbf{p}_0} - K_{\mathbf{p}}\|_{H^1} + \|K_{\mathbf{p}} - \nabla_{\mathbf{p}}\Phi(\mathbf{p})\|_{H^1} \\ &\le M(|\mathbf{p}_0 - \mathbf{p}| + \delta_C) \le M(T_1 + \delta_C). \end{aligned} \tag{2.40}$$

We examine the denominator of (2.28), observing that

$$\left|(E_i(\mathbf{p}_0,\mathbf{p}), \Psi_i^\dagger(\mathbf{p}_0))\right| \le M(T_1 + \delta_C),$$

which for T_1 and δ_C small enough implies the denominator is uniformly bounded away from zero. From (2.28) we then find the bound

$$|\mathbf{p}'| \le M\left(T_1\|W\|_{H^1} + \|W\|_{H^1}^2 + \delta_0(1 + \|W\|_{H^1})\right). \tag{2.41}$$

We combine the estimates above to obtain the bound below on the forcing term $\mathcal{G}$ of (2.30):

(2.42)
$$\|\mathcal{G}\|_{H^1} \leq M\left(|\mathbf{p}_0 - \mathbf{p}| \cdot \|W\|_{H^1} + \|W\|_{H^1}^2 + \delta_0(1 + \|W\|_{H^1})\right)(1 + T_1 + \delta_C).$$

For δ_C and T_1 small enough, we may neglect the last factor on the right-hand side of (2.42). The estimate (2.34) then yields

$$\|\mathcal{G}\|_{H^1} \leq M(T_1 e^{-k(t-t_0)} T_0 + e^{-2k(t-t_0)} T_0^2 + \delta_0(1 + T_0)). \tag{2.43}$$

With this estimate in hand we replace the quantity $\|\mathcal{G}\|_{H^1}$ in (2.35) to obtain

(2.44)
$$\|W(t)\|_{H^1}$$
$$\leq M\left(e^{-k(t-t_0)}\epsilon + \int_{t_0}^{t} e^{-k(t-t')}\left(e^{-K(t'-t_0)} T_1 T_0 + e^{-2k(t'-t_0)} T_0^2 + \delta_0(1 + T_0)\right) dt'\right).$$

Multiply the inequality above by $e^{k(t-t_0)}$, evaluate the integrals, and take the supremum over $t \in [t_0, \tau]$. Then results the inequality

(2.45)
$$T_0(\tau) \leq M_1(\epsilon + (\tau - t_0) T_0(\tau) T_1(\tau) + T_0^2(\tau)(1 - e^{-k(\tau-t_0)}) + \delta_0(1 + T_0(\tau)) e^{k(\tau-t_0)}),$$

valid for some $M_1 > 0$. The inequality above controls T_0 so long as τ is close to t_0 and T_0 is small. We impose two conditions on τ which control, respectively, the second and fourth terms on the right-hand side of (2.45):

$$T_1 \leq \frac{1}{2M_1(\tau - t_0)} \tag{2.46}$$

and

$$\tau - t_0 \leq \frac{\ln(\epsilon/\delta_0)}{k}. \tag{2.47}$$

For $\tau - t_0$ so small that both (2.46) and (2.47) hold we may write (2.45) as

$$T_0(\tau) \leq \frac{2M_1}{1 - 2\epsilon M_1}(2\epsilon + T_0^2(1 - e^{-k(\tau-t_0)})). \tag{2.48}$$

For ϵ small enough, in terms of M_1, this inequality implies that either $T_0 < r_1(\tau)$ or $T_0 > r_2(\tau)$, where $r_1 < r_2$ are the two roots of the equation

$$2\epsilon - \left(\frac{1}{2M_1} - \epsilon\right) r + r^2(1 - e^{-k(\tau-t_0)}) = 0.$$

Since T_0 is continuous, $T_0(t_0) = \epsilon$, and $\lim_{\tau \to t_0^+} r_2(\tau) = \infty$, it follows that initially $T_0(t_0) \leq r_1(t_0)$ and hence by continuity of T_0, r_1, and r_2 with respect to τ we have the inequality $T_0(\tau) \leq r_1(\tau)$, valid so long as both the conditions (2.46) and (2.47) hold. Moreover, the inequality $r_1(\tau) \leq \lim_{\tau \to \infty} r_1(\tau) = M_* \epsilon$ holds for some $M_* > 0$. This permits us to rewrite the bound on T_0 in the form

$$T_0(\tau) \leq M_* \epsilon, \tag{2.49}$$

where M_*, which quantifies the possible secular growth after one renormalization, is independent of $\tau > t_0$.

We now investigate the range of τ for which we may impose the condition (2.46). From the definition (2.33) of T_1 and the estimates (2.41) and (2.39) we find that

$$\begin{aligned} T_1(t) &\leq \int_{t_0}^{t} |\mathbf{p}'(t')| dt' \\ &\leq M \int_{t_0}^{t} \left(T_1 \|W\|_{H^1} + \|W\|_{H^1}^2 + \delta_0 (1 + \|W\|_{H^1}) \right) dt' \\ &\leq M_2 \left(T_1 T_0 + T_0^2 + \delta_0 (1 + T_0)(t - t_0) \right). \end{aligned} \tag{2.50}$$

Isolating T_1 on the left-hand side of the last estimate above yields

$$T_1(t) \leq \frac{M_2 (T_0^2 + \delta_0 (1 + T_0)(t - t_0))}{1 - M_2 T_0}, \tag{2.51}$$

which, in light of (2.49), becomes

$$T_1(t) \leq M \left(\epsilon^2 + \delta_0 (t - t_0) \right). \tag{2.52}$$

The condition (2.47) yields the inequality $\delta_0 (t - t_0) \leq \frac{\delta_0 \ln(\epsilon/\delta_0)}{k}$, which permits us to rewrite (2.52) as

$$T_1(t) \leq M(\epsilon^2 + \delta_0 \ln(\epsilon/\delta_0)), \tag{2.53}$$

valid so long as conditions (2.46) and (2.47) hold and ϵ and δ_0 are small enough. In particular the condition (2.46) can be replaced with the slightly stronger explicit constraint on τ,

$$\tau - t_0 \leq \frac{M}{\epsilon^2 + \delta_0 \ln(\epsilon/\delta_0)}. \tag{2.54}$$

2.2. The renormalization group decompositions.

We may now construct the decompositions $\{(\pi_n, \tau_n)\}_{n=0}^{\infty}$ of the phase space alluded to in the introduction, and employ the corresponding renormalized equations to bound the remainder and track the flow on the manifold. An important tool is the estimate (2.49), valid for τ satisfying the constraints (2.47) and (2.54). For a given $\delta_0 < \epsilon$, there exists $\hat{\epsilon}(\delta_0) > 0$ such that for all $\epsilon \geq \hat{\epsilon}(\delta_0)$ the condition (2.54) is more strict. We call this case the initial transient, and the complimentary case the asymptotic state; see Figure 2.2. We note here that the condition $\epsilon \geq \hat{\epsilon}(\delta_0)$ implies that δ_0 is exponentially small in terms of ϵ, i.e., $\delta_0 \ll e^{-1/\epsilon}$. However, this separation of scales arises naturally in the pulse-pulse interactions we consider in which $\delta_0 = O(e^{-l})$ for some pulse separation l.

2.2.1. Initial transient.

If the initial remainder, W_0, is large enough, i.e., $\|W_0\|_{H^1} = \epsilon > \hat{\epsilon}(\delta_0)$, then the terms in (2.45) arising from the relaxation into a neighborhood of the manifold dominate those due to the forcing terms. The relevant condition on τ for (2.49) to hold is (2.54), which we write as

$$\tau - t_0 \leq \frac{M_3}{\epsilon^2 + \delta_0 \ln(\epsilon/\delta_0)}, \tag{2.55}$$

for some $M_3 > 0$. We may rewrite (2.49) as

$$T_0(\tau) \leq \epsilon M_* \qquad \text{for} \quad \tau - t_0 \leq \frac{M}{\epsilon^2 + \delta_0 \ln(\epsilon/\delta_0)}, \tag{2.56}$$

which is equivalent to

$$\|W(t)\|_{H^1} \leq \epsilon M_* e^{-k(t-t_0)}. \tag{2.57}$$

In particular we obtain the upper bound

$$\|W(t_1)\|_{H^1} \leq \epsilon M_* e^{-kM_3/(\epsilon^2+\delta_0 \ln(\epsilon/\delta_0))} \qquad \text{for} \quad t_1 = t_0 + \frac{M_3}{\epsilon^2 + \delta_0 \ln(\epsilon/\delta_0)}. \tag{2.58}$$

We are in a position to close the estimates we have developed. For $\epsilon > \hat{\epsilon}(\delta_0)$ we have $\epsilon^2 \gg \delta_0 \ln(\epsilon/\delta_0)$ and the coefficient

$$M_* e^{-kM_3/(\epsilon^2+\delta_0 \ln(\epsilon/\delta_0))} \ll 1 \tag{2.59}$$

for ϵ small enough. This expresses the fact that the renormalization period $[t_0, t_1]$ was long enough that the exponential decay of the semigroup can overcome the short-term secular growth implicit in the factor M_*. At the end of the time period $[t_0, t_1]$ we have a remainder of size $\|W(t_1)\|_{H^1}$, which is much smaller than the original remainder $\|W(t_0)\|_{H^1}$.

We iterate the renormalization procedure outlined above, setting $\epsilon_0 = \epsilon$ and defining

$$\epsilon_n = \epsilon_{n-1} M_* e^{-kM_3/(\epsilon_{n-1}^2+\delta_0 \ln(\epsilon_{n-1}/\delta_0))} \tag{2.60}$$

for $n = 1, \ldots, \hat{n}$, where $\hat{n} = \hat{n}(\delta)$ is specified below. We emphasize that $\epsilon_{n+1} \ll \epsilon_n$ when $\epsilon_n > \hat{\epsilon}(\delta_0)$. We renormalize the remainder W at time t_n according to

$$\hat{W}_n = (1/\epsilon_n)\, W(t_n), \tag{2.61}$$

which in light of (2.58) affords the bound $\|\hat{W}_n\|_{H^1} \leq 1$. From Proposition 2.2, so long as $d(\mathbf{p}_{n-1}, \partial\mathcal{K}) > \epsilon_n^{1/2}$, we may find $\mathbf{p}_n \in \mathcal{K}$ such that

$$W_n \equiv \epsilon_n \hat{W}_n + \Phi(\mathbf{p}(t_n)) - \Phi(\mathbf{p}_n) \tag{2.62}$$

lies in $X_{\mathbf{p}_n}$. Moreover, (2.22) implies

$$|\mathbf{p}(t_n) - \mathbf{p}_n| \leq M\epsilon_n |\mathbf{p}_n - \mathbf{p}_{n-1}| \leq M\epsilon_n T_1(t_n^-) \leq M\epsilon_n^3, \tag{2.63}$$

where we employed (2.53) and $\epsilon_n^2 \gg \delta_0 \ln(\epsilon_n/\delta_0)$ in the last inequality. We also have $\|W_n\|_{H^1} \leq M\epsilon_n$ for some constant M independent of n. The evolution equations (2.28) and (2.30) with initial data $\mathbf{p}_n$ and W_n at time $t = t_n$ may be solved on $[t_n, t_{n+1}]$, where

$$t_{n+1} = t_n + \frac{M_3}{\epsilon_n^2 + \delta_0 \ln(\epsilon_n/\delta_0)}, \tag{2.64}$$

with the resulting bounds

$$\|W(t)\|_{H^1} \leq \epsilon_n M_* e^{-k(t-t_n)} \qquad \text{for} \quad t \in [t_n, t_{n+1}], \tag{2.65}$$

valid uniformly in n.

The sequence of renormalization gauges ϵ_n converges rapidly to zero, and after some small number $\hat{n} = \hat{n}(\delta_0)$ of iterations we arrive at the situation $\epsilon_{\hat{n}} < \hat{\epsilon}(\delta)$, and the initial transient is completed. From the recursive definition of ϵ_n it is straightforward to rewrite the bound (2.65) as

$$\|W(t)\|_{H^1} \leq \epsilon M^{n+1} e^{-k(t-t_0)} \qquad \text{for} \quad t \in [t_n, t_{n+1}], \quad n = 0, \dots, \hat{n}, \tag{2.66}$$

where $M > M_*$ may be chosen independently of n; see [25] for details. The term M^{n+1} may be interpreted as a logarithmic correction to the exponential decay rate k associated with the individual linearized operators. This correction arises from the time dependent modulation of the linearized operators induced by the slow flow on the manifold.

2.2.2. Asymptotic state. At the $\hat{n}$th iteration we have $\epsilon_{\hat{n}} \leq \hat{\epsilon}(\delta_0)$, the constraint (2.47) is more exigent than (2.54), and the forcing terms dominate the evolution of the parameters $\mathbf{p}$. We have the evolution equations (2.28) for $\mathbf{p}$ and (2.30) for the remainder W with initial data $\mathbf{p}_{\hat{n}}(0) = \mathbf{p}_{\hat{n}}$ and $W_{\hat{n}}(0)$ given by (2.62) evaluated at $n = \hat{n}$. We iterate the procedure outlined in (2.35)–(2.49), which in light of the constraint (2.47) yields the estimate

$$T_0(\tau) \leq \epsilon_{\hat{n}} M_* \qquad \text{for} \quad \tau - t_{\hat{n}} \leq k^{-1} \ln(\epsilon_{\hat{n}} / \ln \delta_0), \tag{2.67}$$

which is equivalent to

$$\|W(t)\|_{H^1} \leq \epsilon_{\hat{n}} M_* e^{-k(t-t_{\hat{n}})} \qquad \text{for} \quad \tau - t_{\hat{n}} \leq k^{-1} \ln(\epsilon_{\hat{n}}/\delta_0) \tag{2.68}$$

and, in particular,

$$\|W(t_{\hat{n}+1})\|_{H^1} \leq \delta_0 M_* \qquad \text{for} \quad t_{\hat{n}+1} = t_{\hat{n}} + k^{-1} \ln(\epsilon_{\hat{n}}/\delta). \tag{2.69}$$

We define $\epsilon_* = \delta_0 M_*$, setting $\epsilon_n = \epsilon_*$ and $t_{n+1} = t_n + k^{-1} \ln(\epsilon_*/\delta_0) = t_n + k^{-1} \ln(M_*)$ for all $n > \hat{n}$. We note that the time interval $t_{n+1} - t_n$ is precisely that required for the exponential decay to balance the secular growth. This exact balance arises from the precise form of condition (2.47). With an iteration similar to that outlined for the initial transient, we find that

$$\|W(t_n)\|_{H^1} \leq \epsilon_* \qquad \text{for all} \quad n > \hat{n}, \tag{2.70}$$

and, moreover,

$$\|W(t)\|_{H^1} \leq \epsilon_* M_* e^{-k(t-t_n)} \qquad \text{for} \quad t \in [t_n, t_{n+1}], \quad n > \hat{n}, \tag{2.71}$$

where we recall that M_* is independent of δ, ϵ, or n.

We note that the estimate (2.71), taken over the interval $[t_n, t_{n+1}]$, is stronger at the endpoint $t = t_{n+1}$ than at the initial point $t = t_n$. However the variation in the right-hand side is only $O(1)$ and we may rewrite (2.71) as

$$\|W(t)\|_{H^1} \leq M\delta_0 \qquad \text{for} \quad t > t_{\hat{n}+1}, \tag{2.72}$$

for $M = \max\{M_*, M_*^2\}$.

We may now complete the proof of Theorem 2.1. The estimate (2.17) follows from (2.66) and (2.72). The curve $\mathbf{p}(t)$ given by (2.28) is smooth except for the jumps

$\Delta_n \equiv \mathbf{p}_n - \mathbf{p}(t_n)$ at the renormalization times $t = t_n$ which, from Proposition 2.2 and (2.53), satisfy

$$|\Delta_n| \leq \begin{cases} M\epsilon_{n-1}^3 & \text{for} \quad n \leq \hat{n}, \\ M\delta_0^2 & \text{for} \quad n > \hat{n}, \end{cases} \tag{2.73}$$

where M may be chosen independent of n. We replace $\mathbf{p}(t)$ with $\tilde{\mathbf{p}}(t)$, which is a smooth curve verifying $|\mathbf{p} - \tilde{\mathbf{p}}| \leq M(\epsilon e^{-k(t-t_0)} + \delta_0)$ for some M. The remainder W is then replaced with $\tilde{W} = U - \Phi(\tilde{\mathbf{p}})$, which from the smoothness of Φ will also verify the estimate (2.17). The existence of such a smooth curve requires verifying that each jump $|\mathbf{p}(t_n) - \mathbf{p}_n|$ is smaller than the bound $M(\epsilon e^{-k(t_n-t_0)} + \delta_0)$, which follows easily from the estimates at hand for δ_0 and ϵ small enough.

We determine a lower bound on the time T to exit $\mathcal{K}$ by bounding the distance from $\mathbf{p}_0$ to $\mathbf{p}(t)$,

$$|\mathbf{p}_0 - \mathbf{p}(t)| \leq \sum_{i=1}^{n} \left(T_1(t_i^-) + |\Delta_i| \right), \tag{2.74}$$

where $T_1(t_i^-) \equiv \lim_{t \to t_i^-} T_1(t_i)$. Here $n = n(t)$ is determined to be the least integer such that $t < t_n$. From Proposition 2.2 and (2.53) we find that $T_1(t_n^-)$, the accumulated drift in $\mathbf{p}(t)$ over the time interval $[t_{n-1}, t_n]$, satisfies

$$T_1(t_n^-) \leq \begin{cases} M\epsilon_{n-1}^2 & \text{for} \quad n \leq \hat{n}, \\ M\delta_0 & \text{for} \quad n > \hat{n}, \end{cases} \tag{2.75}$$

and hence it dominates the jump $|\Delta_i|$ at the end of the interval. With these estimates in hand, and using the relation $\epsilon_{n+1} \ll \epsilon_n$ for $n < \hat{n}$, the sum (2.74) reduces to

$$|\mathbf{p}_0 - \mathbf{p}(t)| \leq M\left(\epsilon_0^2 + (n - \hat{n})\delta_0\right). \tag{2.76}$$

Finally we observe that $\epsilon_0 = \epsilon$, $|\mathbf{p}_* - \mathbf{p}_0| \leq M\epsilon$, and the intervals $t_{n+1} - t_n = k^{-1}\ln(M_*)$, which allow us to rewrite (2.76) as

$$|\mathbf{p}_* - \mathbf{p}(t)| \leq M\left(\epsilon + (t - t_0)\delta_0\right). \tag{2.77}$$

Assuming that $d(\mathbf{p}_*, \partial\mathcal{K}) = O(1)$, the bound (2.77) readily implies that $\mathbf{p}(t) \in \mathcal{K}$ if $t - t_0 \leq Md(\mathbf{p}_*, \partial\mathcal{K})/\delta_0$ for some $M > 0$. The estimate (2.18) on T then follows for $\mathbf{p}$ and for $\tilde{\mathbf{p}}$.

The evolution for the pulse parameter $\mathbf{p}$ is given by (2.28), which from the estimates established in section 2 may be written in the form

$$p_i' = -\delta_0(\xi, \Psi_i^\dagger(\mathbf{p}))_2 + \delta_0(\xi, \Psi_i^\dagger(\mathbf{p}) - \Psi_i^\dagger(\mathbf{p}_n))_2 + O(T_0^2, T_0T_1, \delta^2) \tag{2.78}$$

for $t \in [t_n, t_{n+1})$. But $|\Psi_i^\dagger(\mathbf{p}) - \Psi_i^\dagger(\mathbf{p}_n)| = O(T_1)$ and after the initial transient, that is, for $t \geq t_{\hat{n}}$ where $\hat{n} = \hat{n}(\delta_0)$, we have $\epsilon_* = O(\delta_0)$ and the inequalities (2.53) and (2.68) imply that $T_0, T_1 = O(\delta)$. In this asymptotic regime, recalling the definition $\xi = \tilde{\xi} + F(\Phi(\mathbf{p}))/\delta_0$, (2.78) reduces to (2.19). Moreover the jumps Δ_n at the times $t = t_n$ are $O(\delta_0^2)$ for $n > \hat{n}$ and are separated by an $O(1)$ time interval. Thus we may choose the smoothed curve $\tilde{\mathbf{p}}$ in such a manner that $|\tilde{\mathbf{p}}' - \mathbf{p}'| = O(\delta_0^2)$ for $t \neq t_i$ for $i = 1, 2, \ldots$. Dropping the tilde notation, we may then write the evolution for the smoothed curve $\mathbf{p}$ as in (2.19). □

3. Applications to PNLS.

We apply the renormalization techniques to the forced PNLS equation

$$i\phi_t + \frac{1}{2}\phi_{xx} \pm |\phi|^2\phi + (i \mp a)\phi - \gamma\phi^* = \delta_0\tilde{\xi}(x,t), \tag{3.1}$$

where $\delta_0 \ll 1$. As a model of pattern formation in the optical parametric oscillator, the parameters $a, \gamma > 0$ represent cavity detuning and pump strength, respectively. The perturbation $\tilde{\xi} \in L^\infty(\mathbf{R}_+, H^1)$ represents the stimulated emission present in the background radiation. Depending upon the nature of the detuning there are two cases: the focusing, which takes the top sign in (3.1), and the defocusing, which takes the bottom sign.

3.1. Stability of fronts in the defocusing PNLS.

For $\gamma \geq 1$, the unforced defocusing PNLS has front solutions

$$\phi = \eta e^{i\theta}\tanh\eta(x-p), \tag{3.2}$$

where

$$\begin{array}{rcl} \eta^2 & = & a + \sqrt{\gamma^2 - 1}, \\ \gamma e^{-2i\theta} & = & -\sqrt{\gamma^2-1} + i. \end{array} \tag{3.3}$$

This solution is linearly stable for $\gamma > 0$ and $a > 0$ [25].

After the change of dependent and independent variables

$$\begin{array}{l} \phi(x) = \eta u(\tilde{x})e^{i\theta}, \\ \tilde{t} = \eta^2 t/2, \\ \tilde{x} = \eta x, \end{array} \tag{3.4}$$

and, dropping the tilde notation, $U = (\Re u, \Im u)^t$ satisfies

$$U_t = \begin{pmatrix} 0 & -(\partial_x^2 - 2|U|^2 + 2\mu) \\ \partial_x^2 - 2|U|^2 + 2 & -\frac{4}{\eta^2} \end{pmatrix} U + \delta_0\Xi, \tag{3.5}$$

where $\mu = \frac{a-\sqrt{\gamma^2-1}}{a+\sqrt{\gamma^2-1}} \leq 1$ and Ξ is derived from the real and imaginary parts of the rescaled $\tilde{\xi}$. This equation readily fits into the form (2.1). The stationary solution ϕ becomes

$$U = \Phi(x-p) = \begin{pmatrix} \tanh(x-p) \\ 0 \end{pmatrix}, \tag{3.6}$$

and the linearization of (3.5) about (3.6) is given by

$$L_p = \begin{pmatrix} 0 & D \\ -C & -\frac{4}{\eta^2} \end{pmatrix}, \tag{3.7}$$

where

$$\begin{array}{rcl} C & = & -(\partial_x^2 + 6\text{sech}^2(x-p) - 4), \\ D & = & -(\partial_x^2 + 2\text{sech}^2(x-p) + 2(\mu - 1)). \end{array} \tag{3.8}$$

From Theorem 3.6 of [25] the spectrum of the operator L_p satisfies

$$\sigma(L_p) = \sigma_s \cup \sigma_0, \tag{3.9}$$

where $\sigma_s \subset \{\lambda \big| \Re\lambda < -k\}$ for some $k > 0$ and $\sigma_0 = \{0\}$ represents a simple eigenvalue of L_p at the origin. This verifies hypothesis (H1). The kernel of L_p is spanned by $\Psi_1 = (-\mathrm{sech}^2(x-p), 0)^t$ with adjoint eigenvector, under the normalization (2.14), given by

$$\Psi_1^\dagger = \begin{cases} \dfrac{-1}{(D^{-1}\mathrm{sech}^2(x), \mathrm{sech}^2(x))_2}\left(D^{-1}\mathrm{sech}^2(x-p), \dfrac{\eta^2}{4}\mathrm{sech}^2(x-p)\right)^t, & \mu \neq \frac{1}{2}, \\ -\dfrac{2}{\pi}\left(\mathrm{sech}(x-p), 0\right)^t, & \mu = \frac{1}{2}. \end{cases}$$

While the operator D is not invertible for $\mu = \frac{1}{2}$, the adjoint eigenvector is bounded for all μ by Lemma 3.4 of [25], verifying hypothesis (H4). We may define the spectral projection π_p by

$$\pi_p U = (U, \Psi_1^\dagger)_2 \Psi_1, \tag{3.10}$$

with kernel $X_p = \pi_p H^1(\mathbf{R})$. As indicated by Proposition 3.2 in the next subsection, the $\mathcal{C}_0$ semigroup S_p generated by L_p verifies hypothesis (H2), where the decay constant k can be taken independent of the position p. Moreover, since Φ is an exact solution of the unforced equation, the kernel of L_p is the translational invariant

$$\Psi_1 = \frac{\partial \Phi}{\partial p}, \tag{3.11}$$

and the compatibility hypothesis (H3) holds with $\delta_C = 0$. We may apply Theorem 2.1 to the front solution (3.2) of the defocusing PNLS.

Theorem 3.1. *Let δ_0 and $\epsilon > 0$ be small enough. Any solution U of the defocusing PNLS equation* (3.5) *corresponding to initial data*

$$U_0 = \Phi(x-p) + W_0$$

with $\|W_0\|_{H^1} \leq \epsilon$ verifies the conclusions of Theorem 2.1 *with $\mathcal{K} = \mathbf{R}$ and $T = \infty$. Moreover, after the initial transient the pulse position p is governed by*

$$p' = -\delta_0(\Xi(t), \Psi_1^\dagger(x-p))_2 + O(\delta_0^2). \tag{3.12}$$

Proof. It only remains to verify that we may take $\mathcal{K} = \mathbf{R}$. Since the parameter p serves only to translate the front, it is straightforward to verify that the hypotheses (H0)–(H4) hold uniformly for $p \in \mathbf{R}$. In particular M and k in (H2) are independent of p, and we may take $\delta_C = 0$ in (H3) for all $p \in \mathbf{R}$. □

3.2. Evolution of pulse trains in the focusing PNLS. In this section we address the evolution of well-separated trains of pulses for the focusing PNLS equation; for simplicity of presentation we neglect the external time dependent forcing term $\tilde{\xi}$. We introduce the new equation parameters

$$\begin{aligned} \eta^2 &= 2(a + \sqrt{\gamma^2 - 1}), \\ \gamma e^{-2i\theta} &= \sqrt{\gamma^2 - 1} + i, \end{aligned} \tag{3.13}$$

and rescale as in (3.4) to obtain the equation below for $U = (\Re u, \Im u)^t$:

$$U_t = \begin{pmatrix} 0 & -(\partial_x^2 + |U|^2 - \mu) \\ \partial_x^2 - |U|^2 - 1 & -\frac{4}{\eta^2} \end{pmatrix} U, \tag{3.14}$$

where $\mu = \frac{a - \sqrt{\gamma^2 - 1}}{a + \sqrt{\gamma^2 - 1}} \leq 1$. This equation fits the form (2.1) for $\mu > 0$, and supports pulse solutions

$$\Phi(x, s) = \begin{pmatrix} \phi(x - s) \\ 0 \end{pmatrix}, \tag{3.15}$$

where $\phi(x) = \sqrt{2}\operatorname{sech} x$, and $s \in \mathbf{R}$ denotes the pulse position.

We examine the evolution of initial data in the neighborhood of the manifold $\mathcal{M} = \left\{\Phi_N(\mathbf{s}) \middle| \mathbf{s} \in \mathcal{K}_l\right\}$, where Φ_N describes N well-separated pulses

$$\Phi_N(x, \mathbf{s}) = \sum_{i=0}^{N} \alpha_i \Phi(x - s_i). \tag{3.16}$$

Here $\mathbf{s} = (s_1, \ldots, s_N)^t \in \mathbf{R}^N$ is the vector of pulse positions and $\alpha_i = \pm 1$ for $i = 1, \ldots, N$ connote a fixed choice of up-down pulse profiles. The set $\mathcal{K}_l$ of admissible pulse trains is given by

$$\mathcal{K}_l = \{\mathbf{s} \in \mathbf{R}^N \big| s_i < s_{i+1} \quad \text{for} \quad i = 1, \ldots, N-1, \qquad \text{and} \qquad \Delta\mathbf{s} \geq l\}, \tag{3.17}$$

where $\Delta\mathbf{s} \equiv \min_{i \neq j} |s_i - s_j|$ and $l > 0$ is a minimum pulse separation. Observe that Φ_N is not an exact stationary solution of (3.14); indeed,

$$\begin{pmatrix} 0 & -(\partial_x^2 + |\Phi_N|^2 - \mu) \\ \partial_x^2 - |\Phi_N|^2 - 1 & -\frac{4}{\eta^2} \end{pmatrix} \Phi_N = F, \tag{3.18}$$

where F is given by

$$F(\mathbf{s}) = -\begin{pmatrix} 0 \\ \left(\sum\limits_{k=1}^{N} \alpha_k \phi_k\right)^3 \end{pmatrix} + \sum_{k=1}^{n} \begin{pmatrix} 0 \\ \alpha_k \phi_k^3 \end{pmatrix}. \tag{3.19}$$

The linearization $L_{\mathbf{s}}$ of (3.1) about $\Phi_N(x, \mathbf{s})$ has the form

$$L_{\mathbf{s}} = \begin{pmatrix} 0 & D \\ -C & -4/\eta^2 \end{pmatrix}, \tag{3.20}$$

where

$$C = -\left(\partial_x^2 - 1 + 3\sum_{i=1}^{N} \phi_i^2 + 6\mathcal{V}\right), \tag{3.21}$$

$$D = -\left(\partial_x^2 - \mu + \sum_{i=1}^{N} \phi_i^2 + 2\mathcal{V}\right), \tag{3.22}$$

with $\phi_i \equiv \phi(x - \mathbf{s}_i)$, and

$$\mathcal{V} = \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \alpha_i \alpha_j \phi_i \phi_j \tag{3.23}$$

represents small cross terms. We decompose the solution U of (3.1) as

$$U = \Phi_N(x, \mathbf{s}) + W(x,t), \tag{3.24}$$

where the evolution for W is given by

$$W_t = L_{\mathbf{s}} W + F + \mathcal{N}(W) + \nabla_{\mathbf{s}} \Phi_N \mathbf{s}', \tag{3.25}$$

with the forcing term F given by (3.19).

3.2.1. Verification of hypothesis. We verify the hypotheses (H0)–(H4) required to apply Theorem 2.1 and derive the modulational equations which govern the slow evolution along the ansatz manifold. It is quite straightforward to verify hypothesis (H0) on the uniform smoothness of the manifold and the forcing term f.

The stability of a single pulse for (3.1) was established in Theorem 3.3 of [25]. Indeed, if L denotes the linearization about a single pulse, then there exists $a_c > 0$ such that for detuning and pump values a and γ satisfying $(a, \gamma) \in [0, a_c) \times (1, \sqrt{1+a^2})$, the spectrum of L consists of a simple eigenvalue at the origin and a remainder which is uniformly bounded in the left-half complex plane. If $L_{\mathbf{s}}$ is the linearization about N well-separated pulses, then the essential spectrum $\sigma_e(L_{\mathbf{s}})$ is determined by the limiting states $\lim_{x\to\pm\infty} \Phi_N$ [16], and thus $\sigma_e(L_{\mathbf{s}}) = \sigma_e(L)$. Moreover, from results of Alexander and Jones [1] (see also [27]) to each localized eigenvalue λ of L, there are associated N eigenvalues $\lambda_1, \ldots, \lambda_N$ of $L_{\mathbf{s}}$, up to multiplicity, such that $\max_{i=1,\ldots,N} |\lambda_i - \lambda|$ decays exponentially with growing pulse separation $l > 0$. In particular we can verify hypothesis (H1), where σ_0 contains the N eigenvalues which approach the origin for large l. These N eigenvalues are the remnants of the translational eigenvalues of the individual pulses.

We verify that $L_{\mathbf{s}}$ satisfies hypothesis (H2) by applying Proposition 4.1 of [25], which for completeness we summarize below.

Proposition 3.2. *Let L be an operator of the form* (3.20) *with the suboperators given by $C = -\partial_x^2 + V_1(x)$ and $D = -\partial_x^2 + V_2(x)$, where V_1 and V_2 are smooth, uniformly bounded potentials. If $\sigma(L)$ satisfies hypothesis* (H1), *then there exist constants $M, k > 0$ such that the associated semigroup $S(t)$ satisfies*

$$\|S(t)u\|_{H^1} \le M e^{-kt} \|u\|_{H^1} \qquad \textit{for all} \;\; t \ge 0, u \in X_1. \tag{3.26}$$

Here the space X_1 is the eigenspace associated with σ_s.

The constants M and k which appear in the estimate (3.26) depend continuously upon $\mathbf{s}$, and thus are uniformly bounded above and below on compact subsets of $\mathcal{K}_l$. However we may formally relax the compactness restriction on $\mathcal{K}_l$ since the individual pulses are stable and the constants M and k are bounded uniformly for large pulse separations.

It remains only to verify (H3)–(H4) about the eigenspace $Y_{\mathbf{s}}$. The location of small eigenvalues of N-pulses has been addressed in detail in [27]. Although we are not linearizing about an exact N-pulse but rather about a quasi-steady train of N-pulses, the aforementioned results can be extended to apply in this case [29]. For completeness we include details of this derivation below. We proceed with a regular expansion of the eigenvectors of σ_0. With tail-tail interactions of pulses $\phi(x - \mathbf{s}_i)$, it is natural to introduce the small quantities $\delta_{ij} = e^{-|\mathbf{s}_i - \mathbf{s}_j|}, \delta = \delta(\mathbf{s}) = e^{-\Delta \mathbf{s}}$, and $\delta_0 = e^{-l}$ which measure the magnitudes of the tail overlap. However the pulse ϕ is degenerate in the sense that it decays at the fast rate $e^{-|x|}$ as $x \to \infty$ associated with the operator C. There is also a slower decay rate $e^{-\sqrt{\mu}|x|}$, associated with the operator

D, which necessitates the introduction of the small quantities $\rho_{ij} = e^{-\sqrt{\mu}|\mathbf{s}_i - \mathbf{s}_j|}$ and $\rho = \rho(\mathbf{s}) = e^{-\sqrt{\mu}\Delta\mathbf{s}}$. Since $\mu < 1$ these small parameters satisfy the relations $\delta_{ij} \leq \delta \leq \delta_0 \ll 1$ and $\delta \ll \rho$.

We introduce the $2 \times N$ matrices

$$\Xi = \begin{pmatrix} \chi_1 & \cdots & \chi_N \\ 0 & \cdots & 0 \end{pmatrix}, \tag{3.27}$$

$$\tilde{\Xi} = \begin{pmatrix} D^{-1}\chi_1 & \cdots & D^{-1}\chi_N \\ \frac{\eta^2}{4}\chi_1 & \cdots & \frac{\eta^2}{4}\chi_N \end{pmatrix}, \tag{3.28}$$

where $\chi_i = \sqrt{2}\operatorname{sech}\tanh(x - s_i)$ is the normalized kernel of $C_i = -(\partial_x^2 + 3\phi_i^2 - 1)$. The uniform invertibility of D is addressed by Proposition A.2 in the appendix. We remark that χ_i corresponds to the translational invariant of the linearization about a single pulse ϕ_i; as such $\dfrac{\partial \Phi_N}{\partial \mathbf{s}_i} = \alpha_i \chi_i$. The columns Ξ_k and $\tilde{\Xi}_k$ of Ξ and $\tilde{\Xi}$ satisfy

$$\begin{array}{rcl} \|L_{\mathbf{s}}\Xi_k\|_{H^1} & \leq & M_0\delta\|\Xi_k\|_{H^1}, \\ \|L_{\mathbf{s}}^\dagger\tilde{\Xi}_k\|_{H^1} & \leq & M_0\delta\|\tilde{\Xi}_k\|_{H^1}. \end{array} \tag{3.29}$$

The eigenvector Ψ_k and associated eigenvalue $\lambda_k \in \sigma_0$ admit an expansion

$$\begin{array}{rcl} \Psi_k & = & \Psi_k^{(0)} + \delta\Psi_k^{(1)} + O(\rho\delta), \\ \lambda_k & = & \delta\lambda_k^{(1)} + O(\rho\delta). \end{array} \tag{3.30}$$

The leading order term $\Psi_k^{(0)}$ takes the form

$$\Psi_k^{(0)} = \Xi\beta_k, \tag{3.31}$$

where $\beta_k = \beta_k^{(0)} + \rho\beta_k^{(1)} + O(\rho^2)$ and $\beta_k^{(0)} = (\beta_{k1}, \ldots, \beta_{kN})^t \in \mathbf{C}^N$ are to be determined from the eigenvalue equation

$$L_{\mathbf{s}}^0\Psi_k = \lambda_k\Psi_k. \tag{3.32}$$

The $O(\delta)$ terms of the eigenvalue equation are

$$L_{\mathbf{s}}\Psi_k^{(1)} + \frac{1}{\delta}L_{\mathbf{s}}\Psi_k^{(0)} = \lambda_k^{(1)}\Psi_k^{(0)}. \tag{3.33}$$

We take the L^2 inner product of the equation above with $\tilde{\Xi}_i$ and use (3.29) to find at leading order in δ

$$\left(\Psi_k^{(0)}, \frac{1}{\delta}L_{\mathbf{s}}^\dagger\tilde{\Xi}_i - \lambda_k^{(1)}\tilde{\Xi}_i\right)_2 = 0, \qquad \text{for} \quad i = 1, \ldots, N. \tag{3.34}$$

Using (3.31) and (3.20), equation (3.34) may be put into the matrix form

$$\left(\frac{4\lambda_k^{(1)}}{\eta^2}\hat{D} + \frac{1}{\delta}\hat{C}\right)\beta_k = 0, \tag{3.35}$$

where the $N \times N$ matrices $\hat{D}$ and $\hat{C}$ have entries

$$\hat{D}_{ij} = (D^{-1}\chi_i, \chi_j)_2, \tag{3.36}$$

$$\hat{C}_{ij} = (C\chi_i, \chi_j)_2. \tag{3.37}$$

Thus $\lambda_k^{(1)}$ is the leading order term in the expansion of the eigenvalue λ_k of $L_{\mathbf{s}}^0$ if and only if $\lambda_k^{(1)}$ is an eigenvalue of the matrix

$$P = -\frac{\eta^2}{4\delta}\hat{D}^{-1}\hat{C}. \tag{3.38}$$

The matrix $\hat{C}$ is self-adjoint, and to leading order has the expression $\hat{C} = 16\hat{C}_0 + O(\delta^2)$, where $\hat{C}_0$ has the tridiagonal form

$$\hat{C}_0 = \begin{pmatrix} d_{11} & \delta_{12} & \cdots & 0 \\ \delta_{12} & d_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \delta_{N-1,N} \\ 0 & \cdots & \delta_{N-1,N} & d_{NN} \end{pmatrix}, \tag{3.39}$$

where

$$d_{ii} = 2\alpha_i(\alpha_{i-1}\delta_{i,i-1} + \alpha_{i+1}\delta_{i,i+1})$$

for $i = 1, \ldots, N$ and $\alpha_0 = \alpha_{N+1} = 0$. To simplify the matrix $\hat{D}$ we begin with Proposition A.2 of the appendix, which yields $D^{-1}\chi_i = D_i^{-1}\chi_i + O(\rho)$. We may calculate directly that

$$(D_i^{-1}\chi_i, \chi_j)_2 = \begin{cases} \theta, & i = j, \\ O(\rho_{ij}), & i \neq j, \end{cases} \tag{3.40}$$

where θ is not zero for any η; see Lemma 3.2 of [25]. These bounds yield the approximation $\hat{D} = \theta I + O(\rho)$, and hence

$$\hat{D}^{-1} = \frac{1}{\theta}I + O(\rho). \tag{3.41}$$

The matrix P takes the form $P = P_0 + O(\rho)$ where $P_0 = -\frac{4\eta^2}{\theta\delta}\hat{C}_0$ is self-adjoint and has a complete set of orthonormal eigenvectors $\beta_k^{(0)}$ and real eigenvalues $\lambda_k^{(1)}$, which give the leading order terms of Ψ_k and λ_k from (3.31). We note that inclusion of higher order terms in the expansion may lead to complex eigenvalues λ_k whose imaginary parts satisfy $\Im\lambda_k = O(\rho\delta)$.

For the adjoint eigenvalue problem

$$L_{\mathbf{s}}^{\dagger}\Psi_k^{\dagger} = \lambda_k^{*}\Psi_k^{\dagger}, \tag{3.42}$$

we take the expansion

$$\tilde{\Psi}_k^{\dagger} = \tilde{\Psi}_k^{\dagger(0)} + \delta\tilde{\Psi}_k^{\dagger(1)} + O(\rho\delta) \tag{3.43}$$

for the unnormalized adjoint eigenvectors whose leading order term is given by

$$\tilde{\Psi}_k^{\dagger(0)} = \tilde{\Xi}\tilde{\beta}_k. \tag{3.44}$$

We take the inner product of (3.42) with Ξ_i, and in a manner similar to the eigenvalue problem, we arrive at the equation

$$P\tilde{\beta}_k = \lambda_k^{(1)}\tilde{\beta}_k, \tag{3.45}$$

with P given by (3.38), and hence $\tilde{\beta}_k = \beta_k$.

We choose the parameters $\mathbf{p}$ for the ansatz Φ_N according to

$$s_i = \alpha_i \sum_{j=1}^{N} \beta_{ji} p_j \tag{3.46}$$

so that the chain rule, (3.31), and (3.30) yield

$$\frac{\partial \Phi_N}{\partial p_j} = \sum_{i=1}^{N} \frac{\partial \Phi_N}{\partial s_i} \frac{\partial s_i}{\partial \mathbf{p}_j} = \sum_{i=1}^{N} (\alpha_i \chi_i)(\alpha_i \beta_{ji}) = \Xi \beta_j = \Psi_j^{(0)} = \Psi_j + O(\delta), \tag{3.47}$$

which verifies the compatibility hypothesis (H3).

To construct the projection $\pi_{\mathbf{p}}$ we determine the multiplicity of the eigenvalues in σ_0. Each eigenvalue in σ_0 has algebraic multiplicity 1 if

$$\theta_i \equiv (\Psi_i, \tilde{\Psi}_i^{\dagger})_2 \neq 0. \tag{3.48}$$

From the expansions (3.31) and (3.44) we find

$$\theta_i = (\Xi \beta_i, \tilde{\Xi} \beta_i)_2 + O(\delta) = \sum_{j=1}^{N} (\Xi_j, \tilde{\Xi}_j)_2 |\beta_{ij}|^2 + O(\delta). \tag{3.49}$$

But from Proposition A.2 and (3.40) we have $(\Xi_j, \tilde{\Xi}_j)_2 = (\chi_j, D_j^{-1} \chi_j) = \theta + O(\rho)$, which is nonzero for ρ small enough. Thus we have

$$\theta_i = \theta |\beta_i|^2 + O(\rho) = \theta + O(\rho), \tag{3.50}$$

and each eigenvalue of σ_0 is simple. Normalizing the adjoint eigenvectors

$$\Psi_i^{\dagger} = \frac{\tilde{\Psi}_i^{\dagger}}{\theta_i},$$

we satisfy the orthogonality condition (2.14). To complete the verification of the stability hypothesis (H4), we observe that so long as the eigenvalues $\lambda \in \sigma_0$ remain simple then the eigenfunctions depend analytically upon the parameters $\mathbf{p} \in \mathcal{K}$. The spectral projection is then given by (2.13).

We simplify the dynamics afforded by (2.28) in the context of the pulse train ansatz for the PNLS equation. Initial data U_0 which lie within ϵ in the H^1 norm of the manifold $\mathcal{M}_l = \left\{ \Phi_N(x, \mathbf{s}) \middle| \Delta \mathbf{s} > l \right\}$ may be decomposed into a manifold parameter $\mathbf{p}_0$ and a remainder W_0 via Proposition 2.2. After the initial transient, the evolution for the pulse parameter $\mathbf{p}$ is given by (2.19) with the forcing term $\xi = \frac{F}{\delta_0}$, where F is given by (3.19). The evolution for the native variables $\mathbf{s}$ is given by

$$s_i' = -\alpha_i \left(F, \sum_{j=1}^{N} \beta_{ji} \Psi_j^{\dagger} \right)_2 + O(\delta^2) = -\alpha_i \left(F, \sum_{j,k=1}^{N} \frac{\beta_{ji} \beta_{jk}}{\theta_k} \tilde{\Xi}_k \right)_2 + O(\delta^2), \tag{3.51}$$

where $\tilde{\Xi}_k$ is the kth column of $\tilde{\Xi}$. Summing first over j in the expression above, and exploiting the approximate orthonormality of the rows of the matrix β, we find

$$s_i' = -\frac{\alpha_i}{\theta_i} (F, \tilde{\Xi}_k)_2 + O(\delta \rho). \tag{3.52}$$

The nonnearest neighbor interactions in the integral above are at most second-order in δ and may be neglected. The leading order terms may be evaluated from the asymptotic relation

$$\int_{\mathbf{R}} \chi_k \phi_k^2 \phi_{k\pm 1} dx = \pm \frac{16}{3} \delta_{k,k\pm 1} + O(\delta^2),$$

where we have used $s_i < s_{i+1}$ for $i = 1, \ldots, N-1$. The evolution equation for the pulse positions $\mathbf{s}_i$ of the ith pulse reduces to

$$s_i' = \frac{4\eta^2}{3\theta} \alpha_i \left(\alpha_{i+1} \delta_{i,i+1} - \alpha_{i-1} \delta_{i,i-1} \right) + O(\delta\rho), \tag{3.53}$$

where for notational convenience we have introduced $\alpha_0 = \alpha_{N+1} = 0$.

For initial data comprised of two pulses $U_0 = \Phi_2 = \alpha_1 \phi(x - s_1) + \alpha_2 \phi(x - s_2)$, there is no initial transient and we obtain directly the leading order slow system for the evolution of the pulse positions

$$s_1' = \frac{4\eta^2}{3\theta} \alpha_1 \alpha_2 \delta_{12}, \tag{3.54}$$

$$s_2' = -\frac{4\eta^2}{3\theta} \alpha_1 \alpha_2 \delta_{12}. \tag{3.55}$$

We find that two pulses of like sign attract and two of opposite sign repel. In particular two pulses of like sign will move together until they are no longer separated by the minimum distance l.

THEOREM 3.3. *Let $(a, \gamma) \in [0, a_c) \times (1, \sqrt{1 + a^2})$ be given and fix $l > 0$ large enough. Then for $\mathbf{s} \in \mathcal{K}_l$ given by* (3.17) *and ϵ small enough, the solution U of the focusing PNLS equation* (3.14) *corresponding to the initial data*

$$U_0 = \Phi_N(\mathbf{s}) + W,$$

with $\|W\|_{H^1} \leq \epsilon$, satisfies the conclusions of Theorem 2.1 *where $\delta_0 = e^{-l}$. Moreover, after the initial transient the pulse positions are governed by the system* (3.53) *for $t \in [t_0, t_0 + T]$. In the case of a pulse train with alternating signs, $\alpha_i \alpha_{i+1} = -1$, the pulse separation increases with time and we may take $T = \infty$.*

4. Discussion. We have shown that manifolds comprised of quasi-stationary solutions of dispersive equations can describe the long-time evolution of nearby orbits; moreover, this description is stable under time dependent perturbations. The ansatz manifolds we have employed are global but not invariant under the flow; however, they attract the flow into a thin neighborhood and the flow restricted to the manifold recovers the salient dynamics. While this result is consistent with the presence of a nearby invariant manifold, the existence of such a manifold is not obvious, particularly if an arbitrary time dependent perturbation $\delta\xi(x, t)$ is added to the equation. There are other techniques which have some similarities to the approach we have taken here; among these we mention that of Kirrmann, Schneider, and Mielke [20] (see also [31]) and Grenier [14, 15], which involve the splitting of a solution into a leading order term and a residual, and then bounding the residual for long time periods.

For the pulse trains of the focusing PNLS equation, we capture the effect of the tail-tail interactions whose magnitude is governed by the separation and the spatial decay rate of the pulses. In the scaling we have chosen the sech pulses decay like $e^{-|x|}$

as $x \to \infty$, and the pulse evolution and size of the residual are on the order of $\delta = e^{-l}$, with l being the pulse separation. However, there is a slower spatial decay rate, $\sqrt{\mu} < 1$, associated with stationary solutions of PNLS, and indeed the application of Theorem 3.3 requires the smallness of $\rho = e^{-l\sqrt{\mu}}$. As $\gamma \to \sqrt{1+a^2}$, the slower rate μ tends to zero and the minimum pulse separation required to apply Theorem 3.3 grows. This effect is seen in numerical simulations for small μ; well-separated like-signed pulses attract but at a μ dependent critical separation the attraction is arrested and a stable two-pulse is formed.

As was observed in [19], there is a connection between the focusing PNLS equation and the evolution of pulses in the phase sensitive amplification (PSA) equation for pulse amplitudes in fiber optical systems [21]. We sketch here an argument which can make this connection rigorous. We write the focusing PNLS equation (3.14) as

$$U_t = \begin{pmatrix} 0 & D(U) \\ -C(U) & -4/\eta^2 \end{pmatrix} U, \tag{4.1}$$

where $C(U) = -(\partial_x^2 + |U|^2 - 1)$ and $D(U) = -(\partial_x^2 + |U|^2 - \mu)$, and $U = (U_1, U_2)^t$. We eliminate U_2 to leading order to find

$$U_{1t} + \frac{\eta^2}{4} D(U_1)C(U_1)U_1 = O(|U_t|^2, |U_{tt}|, |U_{xx}||U_2|^2). \tag{4.2}$$

The left-hand side of (4.2) is exactly the PSA equation in the limit of closely spaced amplifiers ($\sigma = 0$ in (1.2) of [30]). Further analysis based upon the results of section 3.2 shows that, in the asymptotic regime of the pulse train evolution, $\|U_t\|_{H^1} = O(\delta), \|U_{tt}\|_{H^1} = O(\delta^2), \|U\|_{H^2} = O(1)$, and $\|U_2\|_{H^1} = O(\delta)$; we may conclude that the right-hand side of (4.2) is $O(\delta^2)$ in L^2. We apply the renormalization machinery of section 2 and exploit the smoothing properties of the semigroup associated with the fourth-order dissipative operator DC, which demonstrates the equivalence, to leading order, of the pulse train evolution in PNLS and PSA. These results will be presented in full detail elsewhere.

While in some sense the hypotheses (H1)–(H4) may be taken as a definition of a "good ansatz," there are several natural directions to extend the results of section 2. Certainly an investigation of an algebraic dichotomy for the semigroup rather than exponential one implicit in hypotheses (H1)–(H2) is warranted. An interesting case would be an exact ansatz whose essential spectrum lies in the left-half plane and touches the origin, with σ_0 comprised of a single eigenvalue at the origin with finite algebraic multiplicity. It is not clear that an algebraic dichotomy could be exploited in the case of a quasi-stationary ansatz, for which the splitting of the eigenvalues in σ_0 would seem to require the use of an exponentially weighted space to push back the essential spectrum and restore the exponential dichotomy. The stability hypothesis (H4) might be relaxed to permit the structure or the dimension of the projection to change. This could arise, for example, in a pulse-splitting ansatz. The splitting of a pulse would be signaled by the angle θ_i, given by (3.48), becoming zero. The dimension of the manifold would increase, and as one can readily see from (3.53), the pulse dynamics could change substantially.

Appendix A. Invertibility of D.

We consider the invertibility of the operator D given by (3.22). We recall the definition $\rho(\mathbf{s}) = e^{-l\sqrt{\mu}}$.

LEMMA A.1. *For l large enough, the operator D_0 given by*

$$D_0 = -(\partial_x^2 - \mu + P_1 + \cdots + P_N),$$

where $P_i = \phi^2(x - \mathbf{s}_i)$, is uniformly boundedly invertible for all $\mathbf{s} \in \mathcal{K}_l$.

Proof. For each $i = 1, \dots, N$, the operator $D_i = -(\partial_x^2 - \mu + P_i)$ is boundedly invertible (see section 3.2 of [25]). Let $\chi_1, \dots, \chi_N$ be a partition of unity satisfying

$$\mathrm{supp}\,\chi_i \subset \left(\frac{\mathbf{s}_i + \mathbf{s}_{i-1}}{2} - 1, \frac{\mathbf{s}_i + \mathbf{s}_{i+1}}{2} + 1 \right),$$

where we take $\mathbf{s}_0 = -\infty$ and $\mathbf{s}_{N+1} = \infty$. We solve the equation

$$Df = g \tag{A.1}$$

for $g \in L^2$ for f by iteration. Set $g^{(1)} = g$ and for $n = 1, 2, \dots$ define

$$f_i^{(n)} = D_i^{-1} \chi_i g^{(n)}, \tag{A.2}$$

$$f^{(n)} = \sum_{i=1}^N f_i^{(n)}, \tag{A.3}$$

$$g^{(n+1)} = D_0 f^{(n)} - g^{(n)}. \tag{A.4}$$

From this construction it follows that

$$D_0(f^{(1)} + \cdots + f^{(n)}) = g - g^{(n+1)}. \tag{A.5}$$

Note that $\|f_i^{(n)}\|_{H^2} \leq M_0 \|\chi_i g^{(n)}\|_2$ for some $M_0 > 0$ independent of i and n, and in particular

$$\|f^{(n)}\|_{H^2} \leq M_0 \|g^{(n)}\|_2. \tag{A.6}$$

From (A.2) and (A.4) we find

$$g^{(n+1)} = \sum_i \sum_{j \neq i} P_j f_i^{(n)}. \tag{A.7}$$

Moreover, since $\mathrm{supp}\,\chi_i g_i^{(n)} \subset \mathrm{supp}\,\chi_i$ and $f_i^{(n)}$ is uniformly bounded, it follows from (A.2) that $f_i^{(n)}$ decays exponentially outside of $\mathrm{supp}\,\chi_i$; indeed,

$$|f_i^{(n)}(x)| \leq M_1 e^{-\sqrt{\mu}|x - \mathbf{s}_i|} \|f\|_\infty.$$

From this estimate and (A.7) a straightforward calculation shows that

$$\|g^{(n+1)}\|_2 \leq M_2 \rho \|f^{(n)}\|_\infty \leq M_2 \rho \|f^{(n)}\|_{H^2}. \tag{A.8}$$

Thus we find that

$$\|g^{(n+1)}\|_2 \leq (M_2 M_0 \rho)^n \|g\|_2,$$

and

$$\|f^{(n)}\|_{H^2} \leq M_0 (M_2 M_0 \rho)^{n-1} \|g\|_2.$$

From the equality (A.5) it follows that $f = \sum_{n=1}^\infty f^{(n)}$ satisfies (A.1), and

$$\|f\|_{H^2} \leq \frac{M_0}{1 - M_0 M_2 \rho} \|g\|_2,$$

which is finite if l is large enough. □

Proposition A.2. *Let D be given by $D = D_0 + \mathcal{V}$, where $\|\mathcal{V}\|_{*,H^1} = O(\delta)$ is given by* (3.23)*. Then for l large enough, D is boundedly invertible, and, moreover,*

$$\|D^{-1} - D_0^{-1}\|_{*,H^1} = O(\delta). \tag{A.9}$$

In particular, if $g = h\text{sech}(x - \mathbf{s}_i)$ for $h \in L^\infty$, then

$$D^{-1}g = D_i^{-1}g + O(\rho), \tag{A.10}$$

where $D_i = -(\partial_x^2 - \mu + P_i)$.

Proof. Since D is a small perturbation of D_0 we have the identity

$$D^{-1} = (I + D_0^{-1}\mathcal{V})^{-1}D_0^{-1}. \tag{A.11}$$

Moreover $(I + D_0^{-1}\mathcal{V})^{-1} = I + O(\delta)$ and (A.9) holds. The estimate (A.10) follows from observing from (A.2) that

$$f^{(1)} = D_i^{-1}\chi_i g^{(1)} + O(\delta) = D_i^{-1}g + O(\delta),$$

and $\|f^{(n)}\|_{H^1} \leq M\rho^{n-1}$ for some $M > 0$ and all $n \geq 2$. □

Acknowledgments. We are thankful to Todd Kapitula, Bjorn Sandstede, and Gene Wayne for many helpful discussions and to the referees for many helpful comments which led to improvements in the presentation of results.

REFERENCES

[1] J.C. Alexander and C.K.R.T. Jones, *A topological invariant arising in the stability analysis of traveling waves,* J. Reine. Angew. Math., 410 (1990), pp. 167–212.

[2] N. Alexeeva, I. Barashenkov, and D. Pelinovsky, *Dynamics of the parametrically driven NLS solitons beyond the onset of the oscillatory instability,* Nonlinearity, 12 (1999), pp. 103–140.

[3] D. Anderson, *Variational approach to nonlinear pulse propagation in optical fibers,* Phys. Rev. A, 27 (1983), pp. 3135–3144.

[4] P.W. Bates, K. Lu, and C. Zeng, *Existence and persistence of invariant manifolds for semiflows in Banach space,* Mem. Amer. Math. Soc., 135 (1998).

[5] P.W. Bates, K. Lu, and C. Zeng, *Persistence of overflowing manifolds for semiflow,* Comm. Pure Appl. Math., 52 (1999), pp. 983–1046.

[6] P. Bates and C.K.R.T. Jones, *Invariant manifolds for semilinear partial differential equations,* Dyn. Reported, 2 (1989), pp. 1–38.

[7] J. Bona, K. Promislow, and C.E. Wayne, *Higher order asymptotics of decay for nonlinear, dispersive, dissipative wave, equations,* Nonlinearity 8, (1995), pp. 1179–1206.

[8] J. Bricmont, A. Kupiainen, and G. Lin, *Renormalization group and asymptotics of solutions of nonlinear parabolic equations,* Comm. Pure Appl. Math., 47 (1994), pp. 893–922.

[9] J. Bricmont and A. Kupiainen, *Renormalizing partial differential equations,* in Constructive Physics (Palaiseau, 1994), Lecture Notes in Phys., 446, Springer-Verlag, Berlin, 1995, pp. 83–115.

[10] J. Carr, *Applications of Centre Manifold Theory,* Springer-Verlag, New York, 1981.

[11] 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.

[12] T. Gallay, *A center-stable manifold theorem for differential equations in Banach spaces,* Comm. Math. Phys., 152 (1993), pp. 249–268.

[13] L.Y. Chen, N. Goldenfeld, and Y. Oono, *Renormalization group and singular perturbations: Multiple scales, boundary layers, and reductive perturbation theory,* Phys. Rev. E (3), 54 (1996), pp. 376–394.

[14] E. Grenier, *Oscillatory perturbations of the Navier-Stokes equation,* J. Math. Pures Appl. (9), 76 (1997), pp. 477–498.

[15] E. Grenier, C.K.R.T. Jones, V. Zharnitsky, and S.K. Turitsyn, *Stabilizing effect of dispersion management,* Phys. D, 152/153 (2001), pp. 794–817.

[16] D. Henry, *The Geometric Theory of Semilinear Parabolic Equations,* Lecture Notes in Math. 840, Springer-Verlag, Berlin, 1981.

[17] T. Kato, *Quasi-linear equations of evolution, with applications to partial differential equations*, in Spectral Theory and Differential Equations, Lecture Notes in Math. 448, A. Dold and B. Eckmann, eds., Springer-Verlag, New York, pp. 25–70.

[18] W. Kath and N. Smyth, *Soliton evolution and radiation loss for the nonlinear Schrödinger equation,* Phys. Rev. E (3), 51 (1995), pp. 1484–1492.

[19] T. Kapitula and B. Sandstede, *Stability of bright solitary-wave solutions to perturbed nonlinear Schrödinger equations,* Phys. D, 124 (1998), pp. 58–103.

[20] P. Kirrmann, G. Scheinder, and A. Mielke, *The validity of modulational equations for extended systems with cubic nonlinearities,* Proc. Roy. Soc. Edinburgh Sect. A, 122 (1992), pp. 85–91.

[21] J.N. Kutz, C.V. Hile, W.L. Kath, R.D. Li, and P. Kumar, *Pulse propagation in nonlinear optical fiber-lines that employ phase sensitive parametric amplifiers,* J. Opt. Soc. Amer. B Opt. Phys., 11 (1994), pp. 2112–2123.

[22] J.N. Kutz and W.L. Kath, *Stability of pulses in nonlinear optical fibers using phase-sensitive amplifiers,* SIAM J. Appl. Math., 56 (1996), pp. 611–626.

[23] A. Mielke, *Locally invariant manifolds for quasilinear parabolic equations,* Rocky Mountain J. Math., 21 (1991), pp. 707–714.

[24] R. Pego and M. Weinstein, *Asymptotic stability of solitary waves,* Comm. Math. Phys., 164 (1994), pp. 305–349.

[25] K. Promislow and J. N. Kutz, *Bifurcation and asymptotic stability in the large detuning limit of the optical parametric oscillator,* Nonlinearity, 13 (2000), pp. 675–698.

[26] A. Sanchez and A.R. Bishop, *Collective coordinates and length-scale competition in spatially inhomogeneous soliton-bearing equations,* SIAM Rev., 40 (1998), pp. 579–615.

[27] B. Sandstede, *Stability of multiple-pulse solutions,* Trans. Amer. Math. Soc., 350 (1998), pp. 429–472.

[28] B. Sandstede, Weak interaction of pulses, manuscript, 2001.

[29] B. Sandstede, *private communication*, Department of Mathematics, Ohio State University, Columbus, OH, 2001.

[30] B. Sandstede, C.K.R.T. Jones, and J.C. Alexander, *Existence and stability of N-pulses on optical fibers with phase-sensitive amplifiers,* Phys. D, 106 (1997), pp. 167–206.

[31] G. Schneider and C.E. Wayne, *Counter-propagating waves on fluid surfaces and the continuum limit of the Fermi-Pasta-Ulam model*, in Equadiff'99: Proceedings of the International Conference on Differential Equations, B. Fielder, K. Gröger, and J. Sprekels, eds., World Scientific, River Edge, NJ, 2000.

[32] A. Vanderbauwhede and G. Iooss, *Center manifold theory in infinite dimensions,* Dyn. Reported, 1 (1992), pp. 125–163.

[33] C.E. Wayne, *Invariant manifolds for parabolic partial differential equations on unbounded domains,* Arch. Ration. Mech. Anal., 138 (1997), pp. 279–306.

[34] C.A. Pillet and C.E. Wayne, *Invariant manifolds for a class of dispersive, Hamiltonian, partial differential equations,* J. Differential Equations, 141 (1997), pp. 310–326.

[35] M.I. Weinstein, *Modulational stability of ground states of nonlinear Schrödinger equations,* SIAM J. Math. Anal., 16 (1985), pp. 472–491.

[36] G.B. Whitham, *Linear and Nonlinear Waves*, Pure and Applied Mathematics, John Wiley and Sons, New York, 1973.

SIAM J. Math. Anal.
Vol. 33, No. 6, pp. 1483–1493

A UNIQUENESS RESULT FOR THE NAVIER–STOKES EQUATIONS WITH VANISHING VERTICAL VISCOSITY*

DRAGOŞ IFTIMIE†

Abstract. Chemin et al. [M2AN Math. Model. Numer. Anal., 34 (2000), pp. 315–335.] considered the three-dimensional Navier–Stokes equations with vanishing vertical viscosity. Assuming that the initial velocity is square-integrable in the horizontal direction and H^s in the vertical direction, they prove existence of solutions for $s > 1/2$ and uniqueness of solutions for $s > 3/2$. Here, we close the gap between existence and uniqueness, proving uniqueness of solutions for $s > 1/2$. Standard techniques are used.

Key words. Navier–Stokes equations, Sobolev spaces

AMS subject classifications. 35Q30, 35Q35, 76D03, 76D05

PII. S0036141000382126

Introduction. Chemin, Desjardins, Gallagher, and Grenier [2] considered the following anisotropic Navier–Stokes equations:

$$(NS_h)\begin{cases} \partial_t v - \nu(\partial_1^2+\partial_2^2)v - \nu_{_V}\partial_3^2 v + v\cdot\nabla v = -\nabla p \quad \text{for} \quad (t,x)\in(0,\infty)\times\mathbb{R}^3,\\ \qquad\qquad\qquad\qquad\qquad \operatorname{div} v = 0 \quad \text{for} \quad (t,x)\in[0,\infty)\times\mathbb{R}^3,\\ \qquad\qquad\qquad\qquad\qquad v\big|_{t=0} = v_0, \end{cases}$$

where $v(t,\cdot):\mathbb{R}^3\to\mathbb{R}^3$ is an incompressible velocity field, p is the pressure, and the constants $\nu > 0$ and $\nu_{_V} \geq 0$ represent the horizontal and vertical viscosities.

Concerning the physical significance of these equations, we refer to [2] and to the references therein. We will simply say that systems of this type can be found in the theory of rotating fluids and also in the study of the Ekman layers for rotating fluids.

As specified above, the vertical viscosity $\nu_{_V}$ may vanish (or converge to 0). For this reason, the classical theory of the Navier–Stokes equations does not apply. Some L^2 energy estimates still hold for (NS_h), but these are not enough to pass to the limit and obtain a weak solution. The strong solution theory doesn't apply either, unless we work in the framework of hyperbolic symmetric systems by ignoring completely the viscosity terms and requiring a lot of regularity for the initial data. Actually, the only result concerning the situation described to be found in the literature is given by [2, Theorems 2, 3].

Theorem 0.1 (see Chemin et al.). *Let $s > 1/2$ be a real number, and let $v_0 \in \mathrm{H}^{0,s}$ be a divergence-free vector field. Then a positive time T and a solution v of (NS_h) defined on $[0,T]\times\mathbb{R}^3$ exist such that*

$$v \in \mathrm{L}^\infty\big(0,T;\mathrm{H}^{0,s}\big)\cap\mathrm{L}^2\big(0,T;\mathrm{H}^{1,s}\big).$$

Furthermore, there exists a constant c such that if $\|v_0\|_{0,s}$ is less than $c\nu$, then we can choose $T = +\infty$. Finally, this solution is unique, provided that $s > 3/2$.

*Received by the editors December 7, 2000; accepted for publication (in revised form) January 10, 2002; published electronically July 9, 2002.

http://www.siam.org/journals/sima/33-6/38212.html

†IRMAR, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex, France (iftimie@maths.univ-rennes1.fr).

The space $\mathrm{H}^{s,s'}$ is a space with Sobolev regularity H^s in (x_1, x_2) and $\mathrm{H}^{s'}$ in x_3, whose precise definition will be given in section 1.

Let us first make some observations regarding the isotropic Navier–Stokes equations. Critical spaces for the three-dimensional (3D) Navier–Stokes equations are the spaces whose homogeneous norm is invariant under the scaling $f(\cdot) \leftrightarrow \lambda f(\lambda\cdot)$. There is no result on existence and uniqueness of solutions for general initial data in a subcritical space (i.e., a space whose homogeneous norm is invariant under the scaling $f(\cdot) \leftrightarrow \lambda^\alpha f(\lambda\cdot)$ for some $\alpha > 1$). For critical spaces, there are many existence and uniqueness results, starting with the classical result for $\mathrm{H}^{\frac{1}{2}}$ of Fujita and Kato [3] and continuing with Besov spaces, Triebel–Lizorkin spaces, etc. We refer to Cannone [1] for details. Let us just note that a borderline case, that of initial data in BMO^{-1} (divergences of BMO vector fields), was recently proved by Koch and Tataru [6]. Initial data in anisotropic critical spaces were considered by the author in [5]. That work contains an existence and uniqueness result [5, Theorem 3.1] for initial data in a critical anisotropic Besov-type space that contains $\mathrm{H}^{0,s}$ for all $s > 1/2$ (but is contained in $\mathrm{H}^{0,\frac{1}{2}}$). The problem of well-posedness for initial data in $\mathrm{H}^{0,\frac{1}{2}}$ seems to be very difficult for the following two reasons. First, the homogeneous version of $\mathrm{H}^{0,\frac{1}{2}}$ is not well defined since it would require defining $\mathrm{H}^{\frac{1}{2}}$ homogeneous regularity in x_3 and it is well known that the homogeneous space $\mathrm{H}^{\frac{1}{2}}$ is not well defined in dimension 1 (or rather it is not a Banach space). Second, the space $\mathrm{H}^{0,\frac{1}{2}}$ is not included in C^{-1} (see [5, Proposition 4.1]) and it seems to be very difficult to prove existence and uniqueness for initial data which is not C^{-1}. (All the spaces of initial data for which existence and uniqueness of solutions are known are embedded in C^{-1}.)

In accordance with what is observed above, the existence part of Theorem 0.1 is very similar to results known for the isotropic Navier–Stokes equations. The key observation is that, although there is not enough regularity in the vertical direction, the partial derivative ∂_3 is always multiplied by u_3 in the nonlinear term, and the divergence-free condition implies that u_3 has enough vertical regularity. Nevertheless, some technical difficulties persist.

In the result of Chemin et al. there is a gap between the existence result and the uniqueness result. This gap is unexpected, especially since, for the full Navier–Stokes equations, $s > 1/2$ is sufficient to get uniqueness within the framework of anisotropic spaces (see [5, Theorem 3.1]). The aim of this work is to close this gap, proving that uniqueness holds when existence does, i.e., $s > 1/2$.

The gap in the proof of uniqueness given by [2] is due to the term $w_3\partial_3 v$ (w is the difference of two solutions and v is one of the two solutions). Roughly speaking, to estimate this term in $\mathrm{H}^{0,\frac{1}{2}}$ one needs at least $\mathrm{H}^{\frac{1}{2}}$ regularity for $\partial_3 v$ in the vertical direction, that is, $\mathrm{H}^{\frac{3}{2}}$ regularity for v in the vertical direction. This demands, of course, that $s > 3/2$. To overcome this difficulty, we propose to estimate the $\mathrm{H}^{0,-\frac{1}{2}}$ norm of w instead of the $\mathrm{H}^{0,\frac{1}{2}}$ norm. This will require only $\mathrm{H}^{\frac{1}{2}}$ regularity for v in the vertical direction, so the hypothesis $s > 1/2$ will suffice.

This point of view implies some difficulties. First, we will require a product theorem for the anisotropic Sobolev spaces where the regularities in the vertical direction are supercritical for one of the terms and subcritical for the other; see Theorem 1.4. Some product theorems are available but only when both regularities are subcritical (see [5, Theorem 1.1] and [4, Theorem 1.1]) or supercritical (see [2, Lemma 1]).

A second difficulty is to estimate a symmetric term of the type $\int v_3 \partial_3 w \cdot \Lambda_3^{-1} w \, \mathrm{d}x$ (Λ_3 is roughly ∂_3; see the next section for the precise definition). Such a term does not appear when making L^2 estimates instead of $\mathrm{H}^{0,-\frac{1}{2}}$. The estimates for symmetric terms are usually complicated when the indices of regularity are not integers. In the previously cited works, the estimates of this type are long and require dyadic decompositions. We will be able to obtain such an estimate through elementary techniques.

The following theorem completes Theorem 0.1 in the sense that the hypothesis $s > 3/2$ is no longer required to get uniqueness of solutions.

THEOREM 0.2. *Let v and $\widetilde{v}$ be two solutions of (NS_h) on $(0,T)$ belonging to $\mathrm{L}^\infty\big(0,T;\mathrm{H}^{0,s}\big)\cap\mathrm{L}^2\big(0,T;\mathrm{H}^{1,s}\big)$, where $s>1/2$. If v and $\widetilde{v}$ have the same initial data, then $v\equiv\widetilde{v}$.*

Although the regularity invoked in the hypothesis of this theorem is not sufficient by itself to define a trace of the velocity v at time $t=0$, it is a classical observation that v satisfying (NS_h) implies some continuity in time of v, namely $v\in\mathrm{C}^0([0,T];\mathrm{H}^{0,r})$ for all $r<s$ (for a proof, see the remarks before (4)). It therefore makes sense to say that v and $\widetilde{v}$ have the same initial data.

The author is able to prove neither uniqueness nor existence (in the regularity class of Theorem 0.1) in the case $s=1/2$. To this respect, we have nothing to add to the comments made for the isotropic Navier–Stokes equations.

In the following section we introduce notation and prove a new product theorem for anisotropic Sobolev spaces. The last section contains the proof of Theorem 0.2.

1. Notation and preliminary results. In the following, C will denote a constant which may change from one relation to another and which may depend on the different parameters $s, s', \dots$ introduced. The constant K is a universal constant which can also change from one relation to another. Two quantities A and B are said to verify the relation $A\simeq B$ if and only if the ratio A/B stays between two positive constants. We denote by $\langle x\rangle$ the quantity $\langle x\rangle=(1+|x|^2)^{\frac12}$.

DEFINITION 1.1. *For $s,s'\in\mathbb{R}$ we define the anisotropic Sobolev space $\mathrm{H}^{s,s'}$ to be the space of those tempered distributions f which satisfy*

$$\|f\|_{s,s'}\overset{def}{=}\big\|\langle\xi'\rangle^s\langle\xi_3\rangle^{s'}\widehat{f}(\xi)\big\|_{\mathrm{L}^2}<\infty,$$

where $\xi'=(\xi_1,\xi_2)$.

The space $\mathrm{H}^{s,s'}$ endowed with the norm $\|\cdot\|_{s,s'}$ is a Hilbert space.

The partial derivative $\partial/\partial x_j$ is denoted by ∂_j. We denote by Λ_3 the operator $\Lambda_3=\big(1-\partial_3^2\big)^{\frac12}$, that is, the operator of multiplication by $\langle\xi_3\rangle$ in the frequency space. Clearly, Λ_3 is an isometry from $\mathrm{H}^{s,s'}$ to $\mathrm{H}^{s,s'-1}$ for all real numbers s and s'.

When we apply an operator to a vector field, we mean that we apply it to each component of the vector field. The $\mathrm{H}^{s,s'}$ norm of a vector field is the Euclidean norm of the $\mathrm{H}^{s,s'}$ norms of the components. If u, v, and w are three vector fields, then $u\cdot\nabla v$ denotes the vector field $\sum_i u_i\partial_i v$, and $u\cdot\nabla v\cdot w$ denotes the scalar $\sum_{i,j}u_i\,\partial_i v_j\,w_j$.

We will need in the proof of Theorem 0.2 certain interpolation properties of the spaces $\mathrm{H}^{s,s'}$. The following proposition is very easy to prove (see [4, Proposition 1.1]).

PROPOSITION 1.2 (interpolation). *Let $s,t,s',t'\in\mathbb{R}$ and $\alpha\in[0,1]$. If $f\in\mathrm{H}^{s,s'}\cap\mathrm{H}^{t,t'}$, then we have that $f\in\mathrm{H}^{\alpha s+(1-\alpha)t,\alpha s'+(1-\alpha)t'}$ and*

$$\|f\|_{\alpha s+(1-\alpha)t,\alpha s'+(1-\alpha)t'}\le\|f\|_{s,s'}^{\alpha}\|f\|_{t,t'}^{1-\alpha}.$$

The multiplicative properties of the anisotropic Sobolev spaces have been studied in several papers [2, 4, 5, 8, 9]. The following result is proved in [4, Theorem 1.1], valid in the periodic case.

Theorem 1.3. *Let $s,t<1$, $s+t>0$, and $s',t'<1/2$, $s'+t'>0$. If $f\in \mathrm{H}^{s,s'}$ and $g\in\mathrm{H}^{t,t'}$, then $fg\in\mathrm{H}^{s+t-1,s'+t'-1/2}$ and there exists a constant C such that*

$$\|fg\|_{s+t-1,s'+t'-\frac12}\le C\|f\|_{s,s'}\|g\|_{t,t'}.$$

The proof in [4], which uses dyadic decompositions, carries over to the case of the full space. Nevertheless, a more elementary proof can be given, as in Theorem 1.4. For further details, see Remark 3.

Theorem 1.3 is not enough for our purposes. Indeed, the regularity we need is "supercritical" in the vertical direction, i.e., greater than 1/2, a situation which is not covered by Theorem 1.3. The purpose of the following theorem is to deal with this difficulty.

Theorem 1.4. *Let $s,t<1$, $s+t>0$, and $s'>1/2$. If $f\in\mathrm{H}^{s,s'}$ and $g\in\mathrm{H}^{t,-\frac12}$, then $fg\in\mathrm{H}^{s+t-1,-\frac12}$ and there exists a constant C such that*

$$\|fg\|_{s+t-1,-\frac12}\le C\|f\|_{s,s'}\|g\|_{t,-\frac12}.$$

The proof will use the following easy lemma.

Lemma 1.5. *Let $s\in\mathbb{R}$ and $n\in\mathbb{N}^*$. A constant C exists such that*

$$\int_{|x|\le R}\langle x\rangle^s\,\mathrm{d}x\le\begin{cases}C\max(1,\langle R\rangle^{s+n}) & \text{if } s+n\neq 0,\\ \sigma_{n-1}\sqrt2(1+\log\langle R\rangle) & \text{if } s+n=0,\end{cases}$$

where the variable of integration x belongs to $\mathbb{R}^n$ and σ_{n-1} denotes the area of the unit sphere in $\mathbb{R}^n$. Moreover, if s is not large (for instance, if $|s|\le 100$), then the constant C can be chosen of the form $C=\frac{K(n)}{|s+n|}$.

Proof of the lemma. Clearly

$$\int_{|x|\le R}\langle x\rangle^s\,\mathrm{d}x=\sigma_{n-1}\int_0^R\langle r\rangle^s r^{n-1}\,\mathrm{d}r\le\sigma_{n-1}\int_0^R\langle r\rangle^{s+n-1}\,\mathrm{d}r.$$

Since $\frac{1+r}{\sqrt2}\le\langle r\rangle\le 1+r$, we deduce that $\langle r\rangle^{s+n-1}\le(1+r)^{s+n-1}$ if $s+n\ge1$, and $\langle r\rangle^{s+n-1}\le\frac{(1+r)^{s+n-1}}{(\sqrt2)^{s+n-1}}$ if $s+n\le1$. It follows that

$$\int_{|x|\le R}\langle x\rangle^s\,\mathrm{d}x\le\sigma_{n-1}\max(1,2^{\frac{1-n-s}{2}})\int_0^R(1+r)^{s+n-1}\,\mathrm{d}r$$
$$=\begin{cases}\sigma_{n-1}\max(1,2^{\frac{1-n-s}{2}})\frac{(1+R)^{s+n}-1}{s+n} & \text{if } s+n\neq0,\\ \sigma_{n-1}\sqrt2\log(1+R) & \text{if } s+n=0.\end{cases}$$

The conclusion follows by using that $(1+R)^{s+n}\simeq\langle R\rangle^{s+n}$ and $\log(1+R)\le 1+\log\langle R\rangle$. □

Remark 1. In the following we don't need to know the behavior of the constant C of Lemma 1.5 as $|s|\to\infty$. Nevertheless, for the sake of completeness we indicate that the constant C can be chosen of the form $\frac{K(n)}{|s+n|}$ as $|s|\to\infty$, too. The proof of this fact is very easy in the case $n\ge2$, and so we include it here. As in the proof of the lemma, we have for $s+n\neq0$ that

$$\int_{|x|\le R}\langle x\rangle^s\,\mathrm{d}x=\sigma_{n-1}\int_0^R\langle r\rangle^s r^{n-1}\,\mathrm{d}r\le\sigma_{n-1}\int_0^R\langle r\rangle^{s+n-2}r\,\mathrm{d}r$$
$$=\frac{\sigma_{n-1}}{s+n}\int_0^R\frac{\mathrm{d}}{\mathrm{d}r}\big(\langle r\rangle^{s+n}\big)\,\mathrm{d}r=\frac{\sigma_{n-1}}{s+n}\big(\langle R\rangle^{s+n}-1\big)\le\frac{\sigma_{n-1}}{|s+n|}\max\big(1,\langle R\rangle^{s+n}\big).$$

Proof of Theorem 1.4. Let $f \in \mathrm{H}^{s,s'}$ and $g \in \mathrm{H}^{t,-\frac{1}{2}}$. We have to estimate the norm

$$\|fg\|_{s+t-1,-\frac{1}{2}} = (2\pi)^{-3} \|\langle \xi' \rangle^{s+t-1} \langle \xi_3 \rangle^{-\frac{1}{2}} \widehat{f} * \widehat{g}(\xi)\|_{\mathrm{L}^2}.$$

By duality,

$$\begin{aligned}(2\pi)^3 \|fg\|_{s+t-1,-\frac{1}{2}} &= \sup_{\|h\|_{\mathrm{L}^2} \leq 1} \int \langle \xi' \rangle^{s+t-1} \langle \xi_3 \rangle^{-\frac{1}{2}} \widehat{f} * \widehat{g}(\xi) h(\xi)\, \mathrm{d}\xi \\ &= \sup_{\|h\|_{\mathrm{L}^2} \leq 1} \iint \langle \xi' + \eta' \rangle^{s+t-1} \langle \xi_3 + \eta_3 \rangle^{-\frac{1}{2}} \widehat{f}(\xi) \widehat{g}(\eta) h(\xi + \eta)\, \mathrm{d}\xi\, \mathrm{d}\eta.\end{aligned}$$

We can further write

(1)
$$\begin{aligned}(2\pi)^3 \|fg\|_{s+t-1,-\frac{1}{2}} \leq{}& \sup_{\|h\|_{\mathrm{L}^2} \leq 1} \underbrace{\iint_{2|\xi'| \geq |\eta'|} \langle \xi' + \eta' \rangle^{s+t-1} \langle \xi_3 + \eta_3 \rangle^{-\frac{1}{2}} \widehat{f}(\xi) \widehat{g}(\eta) h(\xi + \eta)\, \mathrm{d}\xi\, \mathrm{d}\eta}_{I_1} \\ &+ \sup_{\|h\|_{\mathrm{L}^2} \leq 1} \underbrace{\iint_{2|\xi'| \leq |\eta'|} \langle \xi' + \eta' \rangle^{s+t-1} \langle \xi_3 + \eta_3 \rangle^{-\frac{1}{2}} \widehat{f}(\xi) \widehat{g}(\eta) h(\xi + \eta)\, \mathrm{d}\xi\, \mathrm{d}\eta}_{I_2}.\end{aligned}$$

In what follows, whenever we study the dependence of constants on s', we will assume that s' stays bounded (for instance, $s' \leq 100$ or any other universal constant).

We now write I_1 under the form

$$I_1 = \iint_{2|\xi'| \geq |\eta'|} \langle \xi_3 + \eta_3 \rangle^{-\frac{1}{2}} \frac{\langle \xi' + \eta' \rangle^{s+t-1}}{\langle \eta' \rangle^t} \widehat{f}(\xi) \langle \eta' \rangle^t \widehat{g}(\eta) h(\xi + \eta)\, \mathrm{d}\xi\, \mathrm{d}\eta,$$

and we apply Hölder's inequality in the variables ξ' and η' to obtain that

$$\begin{aligned}I_1 \leq \iint \langle \xi_3 + \eta_3 \rangle^{-\frac{1}{2}} \Bigg(& \underbrace{\iint_{2|\xi'| \geq |\eta'|} \frac{\langle \xi' + \eta' \rangle^{2(s+t-1)}}{\langle \eta' \rangle^{2t}} |\widehat{f}(\xi)|^2\, \mathrm{d}\xi'\, \mathrm{d}\eta'}_{I_3} \\ & \times \iint \langle \eta' \rangle^{2t} |\widehat{g}(\eta)|^2 |h(\xi + \eta)|^2\, \mathrm{d}\xi'\, \mathrm{d}\eta' \Bigg)^{\frac{1}{2}} \mathrm{d}\xi_3\, \mathrm{d}\eta_3.\end{aligned}$$

To estimate I_3, we first integrate with respect to η' and then decompose

$$\begin{aligned}\int_{2|\xi'| \geq |\eta'|} \frac{\langle \xi' + \eta' \rangle^{2(s+t-1)}}{\langle \eta' \rangle^{2t}}\, \mathrm{d}\eta' = & \int_{|\eta'| \leq |\xi'|/2} \frac{\langle \xi' + \eta' \rangle^{2(s+t-1)}}{\langle \eta' \rangle^{2t}}\, \mathrm{d}\eta' \\ & + \int_{|\xi'|/2 \leq |\eta'| \leq 2|\xi'|} \frac{\langle \xi' + \eta' \rangle^{2(s+t-1)}}{\langle \eta' \rangle^{2t}}\, \mathrm{d}\eta'.\end{aligned}$$

If $|\eta'| \leq |\xi'|/2$, then $\langle \xi' + \eta' \rangle \simeq \langle \xi' \rangle$. If $|\xi'|/2 \leq |\eta'| \leq 2|\xi'|$, then $\langle \eta' \rangle \simeq \langle \xi' \rangle$. We deduce that

$$\int_{|\eta'| \leq |\xi'|/2} \frac{\langle \xi' + \eta' \rangle^{2(s+t-1)}}{\langle \eta' \rangle^{2t}}\, \mathrm{d}\eta' \simeq \langle \xi' \rangle^{2(s+t-1)} \int_{|\eta'| \leq |\xi'|/2} \frac{1}{\langle \eta' \rangle^{2t}}\, \mathrm{d}\eta' \leq \frac{K}{1-t} \langle \xi' \rangle^{2s}$$

and that

$$\int_{|\xi'|/2\le|\eta'|\le 2|\xi'|}\frac{\langle\xi'+\eta'\rangle^{2(s+t-1)}}{\langle\eta'\rangle^{2t}}\,\mathrm{d}\eta' \simeq \frac{1}{\langle\xi'\rangle^{2t}}\int_{|\xi'|/2\le|\eta'|\le 2|\xi'|}\langle\xi'+\eta'\rangle^{2(s+t-1)}\,\mathrm{d}\eta'$$
$$\le \frac{K}{\langle\xi'\rangle^{2t}}\int_{|\zeta|\le 3|\xi'|}\langle\zeta\rangle^{2(s+t-1)}\,\mathrm{d}\zeta \le \frac{K}{s+t}\langle\xi'\rangle^{2s},$$

where we have used Lemma 1.5 and the change of variables $\zeta=\xi'+\eta'$.

According to the definition of I_3, we obtain from the previous relations that

$$I_3\le C\int\langle\xi'\rangle^{2s}|\widehat{f}(\xi)|^2\,\mathrm{d}\xi',$$

which yields the following estimate for I_1:

$$I_1\le C\iint\left(\int\langle\xi'\rangle^{2s}\langle\xi_3\rangle^{2s'}|\widehat{f}(\xi)|^2\,\mathrm{d}\xi'\int|h(\zeta,\xi_3+\eta_3)|^2\,\mathrm{d}\zeta\right)^{\frac12}$$
$$\times\left(\langle\xi_3+\eta_3\rangle^{-1}\langle\xi_3\rangle^{-2s'}\int\langle\eta'\rangle^{2t}|\widehat{g}(\eta)|^2\,\mathrm{d}\eta'\right)^{\frac12}\mathrm{d}\xi_3\,\mathrm{d}\eta_3.$$

Hölder's inequality applied in the variable (ξ_3,η_3) now gives that

$$I_1\le C\|f\|_{s,s'}\|h\|_{\mathrm{L}^2}\left(\int\langle\eta'\rangle^{2t}\varphi(\eta_3)|\widehat{g}(\eta)|^2\,\mathrm{d}\eta\right)^{\frac12},\tag{2}$$

where

$$\varphi(\eta_3)=\int\frac{1}{\langle\xi_3+\eta_3\rangle\langle\xi_3\rangle^{2s'}}\,\mathrm{d}\xi_3.$$

To estimate φ, we proceed as for I_3 by investigating several pieces and using Lemma 1.5:

$$\int_{|\xi_3|\ge 2|\eta_3|}\frac{1}{\langle\xi_3+\eta_3\rangle\langle\xi_3\rangle^{2s'}}\,\mathrm{d}\xi_3\simeq\int_{2|\eta_3|}^{\infty}\frac{1}{\langle\xi_3\rangle^{2s'+1}}\,\mathrm{d}\xi_3$$
$$\simeq\int_{2|\eta_3|}^{\infty}\frac{1}{(1+\xi_3)^{2s'+1}}\,\mathrm{d}\xi_3\le\frac{K}{\langle\eta_3\rangle^{2s'}}\le\frac{K}{\langle\eta_3\rangle},$$
$$\int_{|\xi_3|\le|\eta_3|/2}\frac{1}{\langle\xi_3+\eta_3\rangle\langle\xi_3\rangle^{2s'}}\,\mathrm{d}\xi_3\simeq\frac{1}{\langle\eta_3\rangle}\int_{|\xi_3|\le|\eta_3|/2}\frac{1}{\langle\xi_3\rangle^{2s'}}\,\mathrm{d}\xi_3\le\frac{K}{(s'-\frac12)\langle\eta_3\rangle},$$
$$\int_{|\eta_3|/2\le|\xi_3|\le 2|\eta_3|}\frac{1}{\langle\xi_3+\eta_3\rangle\langle\xi_3\rangle^{2s'}}\,\mathrm{d}\xi_3\simeq\frac{1}{\langle\eta_3\rangle^{2s'}}\int_{|\eta_3|/2\le|\xi_3|\le 2|\eta_3|}\frac{1}{\langle\xi_3+\eta_3\rangle}\,\mathrm{d}\xi_3$$
$$\le\frac{K}{\langle\eta_3\rangle^{2s'}}\int_{|\zeta|\le 3|\eta_3|}\frac{1}{\langle\zeta\rangle}\,\mathrm{d}\zeta$$
$$\le\frac{K}{\langle\eta_3\rangle^{2s'}}(1+\log\langle\eta_3\rangle)\le\frac{K}{(s'-\frac12)\langle\eta_3\rangle},$$

where we have used in the last relation that $\log\alpha\le\frac{\alpha^{\varepsilon}}{e\varepsilon}$ for all $\alpha\ge 1$ and $\varepsilon>0$. We deduce from the previous relations that

$$\varphi(\eta_3)\le C\langle\eta_3\rangle^{-1},$$

which, plugged into (2), yields the estimate

$$I_1 \leq C\|f\|_{s,s'}\|g\|_{t,-\frac{1}{2}}\|h\|_{\mathrm{L}^2}. \tag{3}$$

To complete the proof, it remains to estimate I_2 (defined in relation (1)). By Hölder's inequality,

$$I_2 \leq (J_1 J_2)^{\frac{1}{2}},$$

where

$$J_1 = \iint \langle\xi'\rangle^{2s}\langle\xi_3\rangle^{2s'}|\widehat{f}(\xi)|^2|h(\xi+\eta)|^2\,\mathrm{d}\xi\,\mathrm{d}\eta$$

and

$$J_2 = \iint_{2|\xi'|\leq|\eta'|} \frac{\langle\xi'+\eta'\rangle^{2(s+t-1)}}{\langle\xi'\rangle^{2s}}\frac{\langle\xi_3+\eta_3\rangle^{-1}}{\langle\xi_3\rangle^{2s'}}|\widehat{g}(\eta)|^2\,\mathrm{d}\xi\,\mathrm{d}\eta.$$

Clearly, $J_1 = \|h\|^2_{\mathrm{L}^2}\|f\|^2_{s,s'}$. To estimate J_2, we first integrate in ξ' and ξ_3. As in the estimate for I_3,

$$\int_{2|\xi'|\leq|\eta'|}\frac{\langle\xi'+\eta'\rangle^{2(s+t-1)}}{\langle\xi'\rangle^{2s}}\,\mathrm{d}\xi' \simeq \langle\eta'\rangle^{2(s+t-1)}\int_{2|\xi'|\leq|\eta'|}\frac{1}{\langle\xi'\rangle^{2s}}\,\mathrm{d}\xi' \leq \frac{K}{1-s}\langle\eta'\rangle^{2t}.$$

Therefore, again using the bound for φ, one deduces that

$$J_2 \leq C\|g\|^2_{t,-\frac{1}{2}}.$$

We conclude that

$$I_2 \leq C\|f\|_{s,s'}\|g\|_{t,-\frac{1}{2}}\|h\|_{\mathrm{L}^2},$$

which, combined with relations (1) and (3), completes the proof of Theorem 1.4. □

Remark 2. It might be useful to know how the constant C in the statement of Theorem 1.4 depends on s, s', and t. Actually, tracking the constants in the proof, the constant C is of the form

$$C = K\left(\frac{1}{\sqrt{1-s}}+\frac{1}{\sqrt{1-t}}+\frac{1}{\sqrt{s+t}}\right)\frac{1}{\sqrt{s'-1/2}},$$

where we have assumed that s' stays bounded (say, $s' \leq 100$).

Remark 3. Theorem 1.4 is a special case of a more general theorem. More precisely, instead of considering $g \in \mathrm{H}^{t,-\frac{1}{2}}$, one may consider $g \in \mathrm{H}^{t,t'}$, where t' verifies $t' \leq s'$ and $s'+t' > 0$. The conclusion is then that $fg \in \mathrm{H}^{s+t-1,t'}$. We chose to prove the special case $t' = -1/2$, sufficient for our purposes, because the proof is considerably simpler. The proof in the general case does not involve any new ideas. In fact, the complication in the proof comes from the fact that the decomposition in ξ' given in relation (1) has to be done in the variable ξ_3 also; therefore one has to examine four pieces instead of just two, but the techniques are identical. Finally, let us note that if we add the hypothesis $t' < s'-1/2$, then the additional decomposition in ξ_3 is not necessary and the proof given here carries over with no modification other than the replacement of $-1/2$ by t'.

2. Proof of the main theorem. We can assume without loss of generality that $s<1$. We will prove that the $\mathrm{H}^{0,-\frac{1}{2}}$ norm of $w=v-\widetilde{v}$ vanishes. In order to estimate $\|w\|_{0,-\frac{1}{2}}$, let us prove that the regularity available is enough to allow us to multiply the equation for $v-\widetilde{v}$ by $\Lambda_3^{-1}w$. First, note that we can write

$$v\cdot\nabla v=\sum_i \partial_i(v_i v).$$

By interpolation and by hypothesis, one has that $v\in\mathrm{L}^4(0,T;\mathrm{H}^{\frac{1}{2},s})$ (see relation (16)). The product theorem 1.3 easily implies that $v_i v\in\mathrm{L}^2(0,T;\mathrm{H}^{0,1-s})$ so $v\cdot\nabla v\in\mathrm{L}^2(0,T;\mathrm{H}^{-1,-s})\subset\mathrm{L}^2(0,T;\mathrm{H}^{-1,-\frac{3}{2}})$. Clearly, $\nu(\partial_1^2+\partial_2^2)v+\nu_{_V}\partial_3^2 v\in\mathrm{L}^2(0,T;\mathrm{H}^{-1,s})+\mathrm{L}^2(0,T;\mathrm{H}^{1,s-2})\subset\mathrm{L}^2(0,T;\mathrm{H}^{-1,-\frac{3}{2}})$. From the equation (NS_h) it follows that $\partial_t v\in\mathrm{L}^2(0,T;\mathrm{H}^{-1,-\frac{3}{2}})$. We deduce that every term in the equations for v and $\widetilde{v}$ belongs to $\mathrm{L}^2(0,T;\mathrm{H}^{-1,-\frac{3}{2}})$ and can therefore be multiplied by $\Lambda_3^{-1}w$, which belongs to $\mathrm{L}^2(0,T;\mathrm{H}^{1,1+s})\subset\mathrm{L}^2(0,T;\mathrm{H}^{1,\frac{3}{2}})$.

Note that the fact that $\partial_t v\in\mathrm{L}^2(0,T;\mathrm{H}^{-1,-\frac{3}{2}})$ and $v\in\mathrm{L}^2(0,T;\mathrm{H}^{1,s})$ implies, by the interpolation theory developed by Lions and Magenes [7, Chapter 1], that $v\in\mathrm{C}^0([0,T];\mathrm{H}^{0,\frac{2s-3}{4}})$. The interpolation property stated in Proposition 1.2 along with the fact that $v\in\mathrm{L}^\infty(0,T;\mathrm{H}^{0,s})$ imply in a classical manner that $v\in\mathrm{C}^0([0,T];\mathrm{H}^{0,r})$ for all $r<s$.

Multiplying the equation for $v-\widetilde{v}$ by $\Lambda_3^{-1}w$, integrating on $(\varepsilon,t)\times\mathbb{R}^3$, letting $\varepsilon\to 0$, and using the continuity in time of $\|w\|_{0,-\frac{1}{2}}$ yields

$$\begin{aligned}\|w(t)\|^2_{0,-\frac{1}{2}}+2\nu\int_0^t\big(\|\partial_1 w(\tau)\|^2_{0,-\frac{1}{2}}+\|\partial_2 w(\tau)\|^2_{0,-\frac{1}{2}}\big)\,\mathrm{d}\tau+2\nu_{_V}\int_0^t\|\partial_3 w(\tau)\|^2_{0,-\frac{1}{2}}\,\mathrm{d}\tau\\ =-2\int_0^t\int v(\tau,x)\cdot\nabla w(\tau,x)\cdot\Lambda_3^{-1}w(\tau,x)\,\mathrm{d}\tau\,\mathrm{d}x\\ -2\int_0^t\int w(\tau,x)\cdot\nabla\widetilde{v}(\tau,x)\cdot\Lambda_3^{-1}w(\tau,x)\,\mathrm{d}\tau\,\mathrm{d}x.\end{aligned}\tag{4}$$

To simplify the notation, we will write v instead of $v(\tau,x)$ and so on. We consider τ fixed, and we evaluate

$$\int v\cdot\nabla w\cdot\Lambda_3^{-1}w\,\mathrm{d}x=\underbrace{\int(v_1\partial_1 w+v_2\partial_2 w)\cdot\Lambda_3^{-1}w\,\mathrm{d}x}_{L_1}+\underbrace{\int v_3\partial_3 w\cdot\Lambda_3^{-1}w\,\mathrm{d}x}_{L_2}\tag{5}$$

and

$$\int w\cdot\nabla\widetilde{v}\cdot\Lambda_3^{-1}w\,\mathrm{d}x=\underbrace{\int(w_1\partial_1\widetilde{v}+w_2\partial_2\widetilde{v})\cdot\Lambda_3^{-1}w\,\mathrm{d}x}_{L_3}+\underbrace{\int w_3\partial_3\widetilde{v}\cdot\Lambda_3^{-1}w\,\mathrm{d}x.}_{L_4}\tag{6}$$

We will now estimate each of these integrals.

Estimate of L_1. According to the product theorem 1.4, one can bound L_1 as follows:

$$|L_1|\le\|v_1\partial_1 w+v_2\partial_2 w\|_{-\frac{1}{2},-\frac{1}{2}}\|\Lambda_3^{-1}w\|_{\frac{1}{2},\frac{1}{2}}\le C\|v\|_{\frac{1}{2},s}\|w\|_{1,-\frac{1}{2}}\|w\|_{\frac{1}{2},-\frac{1}{2}}.$$

By the interpolation property given in Proposition 1.2, one has that

$$\|w\|_{\frac{1}{2},-\frac{1}{2}} \leq \|w\|_{0,-\frac{1}{2}}^{\frac{1}{2}} \|w\|_{1,-\frac{1}{2}}^{\frac{1}{2}}, \tag{7}$$

which leads to

$$|L_1| \leq C\|v\|_{\frac{1}{2},s} \|w\|_{0,-\frac{1}{2}}^{\frac{1}{2}} \|w\|_{1,-\frac{1}{2}}^{\frac{3}{2}}. \tag{8}$$

Estimate of L_3. Again by the product theorem 1.4, we have that

$$|L_3| \leq \|w_1\partial_1\widetilde{v} + w_2\partial_2\widetilde{v}\|_{-\frac{3}{4},-\frac{1}{2}} \|\Lambda_3^{-1}w\|_{\frac{3}{4},\frac{1}{2}} \leq C\|\widetilde{v}\|_{\frac{1}{2},s}\|w\|_{\frac{3}{4},-\frac{1}{2}}^2.$$

By interpolation,

$$\|w\|_{\frac{3}{4},-\frac{1}{2}} \leq \|w\|_{0,-\frac{1}{2}}^{\frac{1}{4}} \|w\|_{1,-\frac{1}{2}}^{\frac{3}{4}},$$

so that

$$|L_3| \leq C\|\widetilde{v}\|_{\frac{1}{2},s} \|w\|_{0,-\frac{1}{2}}^{\frac{1}{2}} \|w\|_{1,-\frac{1}{2}}^{\frac{3}{2}}. \tag{9}$$

Estimate of L_4. We proceed by using Theorem 1.3:

$$\begin{aligned} |L_4| \leq \|w_3\partial_3\widetilde{v}\|_{-\frac{1}{2},\frac{2s-3}{4}} \|\Lambda_3^{-1}w\|_{\frac{1}{2},\frac{3-2s}{4}} &\leq C\|w_3\|_{0,\frac{3-2s}{4}} \|\partial_3\widetilde{v}\|_{\frac{1}{2},s-1} \|w\|_{\frac{1}{2},\frac{-2s-1}{4}} \\ &\leq C\|\widetilde{v}\|_{\frac{1}{2},s} \|w_3\|_{0,\frac{1}{2}} \|w\|_{\frac{1}{2},-\frac{1}{2}}. \end{aligned}$$

But it is trivial to see that

$$\|f\|_{s,s'} = \big(\|f\|_{s,s'-1}^2 + \|\partial_3 f\|_{s,s'-1}^2\big)^{\frac{1}{2}} \leq \|f\|_{s,s'-1} + \|\partial_3 f\|_{s,s'-1}.$$

Therefore, because w is divergence free,

$$\|w_3\|_{0,\frac{1}{2}} \leq \|w\|_{0,-\frac{1}{2}} + \|\partial_3 w_3\|_{0,-\frac{1}{2}} = \|w\|_{0,-\frac{1}{2}} + \|\partial_1 w_1 + \partial_2 w_2\|_{0,-\frac{1}{2}} \leq \|w\|_{0,-\frac{1}{2}} + 2\|w\|_{1,-\frac{1}{2}}.$$

Also using relation (7), we infer that

$$|L_4| \leq C\|\widetilde{v}\|_{\frac{1}{2},s}\big(\|w\|_{0,-\frac{1}{2}}^{\frac{3}{2}} \|w\|_{1,-\frac{1}{2}}^{\frac{1}{2}} + \|w\|_{0,-\frac{1}{2}}^{\frac{1}{2}} \|w\|_{1,-\frac{1}{2}}^{\frac{3}{2}}\big). \tag{10}$$

Estimate of L_2. The proof for L_2 is more delicate. It is a commutator-type estimate and requires an integration by parts. Applying Parseval's formula gives

$$\begin{aligned} L_2 &= \int v_3\partial_3 w \cdot \Lambda_3^{-1} w \,\mathrm{d}x \\ &= (2\pi)^{-3} \int \widehat{v_3\partial_3 w}(\xi) \cdot \widehat{\Lambda_3^{-1} w}(-\xi)\,\mathrm{d}\xi \\ &= (2\pi)^{-6} \int \frac{1}{\langle \xi_3 \rangle} \widehat{v}_3 * \widehat{\partial_3 w}(\xi) \cdot \widehat{w}(-\xi)\,\mathrm{d}\xi \\ &= i(2\pi)^{-6} \iint \frac{\eta_3}{\langle \xi_3 \rangle} \widehat{v}_3(\xi-\eta)\widehat{w}(\eta)\cdot\widehat{w}(-\xi)\,\mathrm{d}\xi\,\mathrm{d}\eta. \end{aligned} \tag{11}$$

Using the change of variables $(\xi,\eta) \leftrightarrow (-\eta,-\xi)$, one can write

$$L_2 = \frac{i}{2}(2\pi)^{-6} \iint \left(\frac{\eta_3}{\langle \xi_3 \rangle} - \frac{\xi_3}{\langle \eta_3 \rangle}\right) \widehat{v}_3(\xi-\eta)\widehat{w}(\eta)\cdot\widehat{w}(-\xi)\,\mathrm{d}\xi\,\mathrm{d}\eta. \tag{12}$$

Now, for $x, y \in \mathbb{R}$, one can check the following identity:

$$\frac{x}{\langle y\rangle} - \frac{y}{\langle x\rangle} = \frac{x-y}{\langle y\rangle} + \frac{(x-y)y(x+y)}{\langle x\rangle\langle y\rangle(\langle x\rangle + \langle y\rangle)}.$$

As $|y| < \langle y\rangle$ and $|x+y| < \langle x\rangle + \langle y\rangle$, we infer that

$$\left|\frac{x}{\langle y\rangle} - \frac{y}{\langle x\rangle}\right| \le |x-y|\left(\frac{1}{\langle x\rangle} + \frac{1}{\langle y\rangle}\right).$$

Therefore, we obtain from (12) that

$$|L_2| \le \frac{1}{2}(2\pi)^{-6}\sum_j \iint |\xi_3 - \eta_3|\left(\frac{1}{\langle \xi_3\rangle} + \frac{1}{\langle \eta_3\rangle}\right)|\widehat{v}_3(\xi-\eta)|\,|\widehat{w_j}(\eta)|\,|\widehat{w}_j(-\xi)|\,\mathrm{d}\xi\,\mathrm{d}\eta.$$

Using again the change of variables $(\xi, \eta) \leftrightarrow (-\eta, -\xi)$, we deduce

$$|L_2| \le (2\pi)^{-6}\sum_j \iint \frac{|\xi_3-\eta_3|}{\langle \xi_3\rangle}|\widehat{v}_3(\xi-\eta)|\,|\widehat{w_j}(\eta)|\,|\widehat{w}_j(-\xi)|\,\mathrm{d}\xi\,\mathrm{d}\eta.$$

As v is divergence free, one has that $\xi_3\widehat{v}_3(\xi) = -\xi_1\widehat{v}_1(\xi) - \xi_2\widehat{v}_2(\xi)$, so $|\xi_3|\,|\widehat{v}_3(\xi)| \le |\xi_1|\,|\widehat{v}_1(\xi)| + |\xi_2|\,|\widehat{v}_2(\xi)|$. It follows that

(13)
$$|L_2| \le (2\pi)^{-6}\sum_j \iint \frac{|\xi_1-\eta_1|\,|\widehat{v}_1(\xi-\eta)| + |\xi_2-\eta_2|\,|\widehat{v}_2(\xi-\eta)|}{\langle \xi_3\rangle}|\widehat{w_j}(\eta)|\,|\widehat{w}_j(-\xi)|\,\mathrm{d}\xi\,\mathrm{d}\eta.$$

Let V be the vector field whose components verify

$$\widehat{V}_j = |\widehat{v}_j|.$$

Obviously, $\|V_j\|_{r,r'} = \|v_j\|_{r,r'}$ for all r, r', and j. We define in the same manner the vector field W. Using the reversed argument of (11), we observe that relation (13) is equivalent to

$$|L_2| \le \int (|D_1|V_1 + |D_2|V_2)W \cdot \Lambda_3^{-1}W\,\mathrm{d}x,$$

where $|D_j|$ denotes the operator of multiplication in the frequency space by $|\xi_j|$. As $|D_j|V_j$ and $\partial_j V_j$ have the same $\mathrm{H}^{r,r'}$ norm for all r, r', and j, the same argument as in the estimate of L_3 shows that

$$\begin{aligned}(14)\qquad |L_2| &\le \int (|D_1|V_1 + |D_2|V_2)W \cdot \Lambda_3^{-1}W\,\mathrm{d}x \le C\|V\|_{\frac{1}{2},s}\|W\|_{0,-\frac{1}{2}}^{\frac{1}{2}}\|W\|_{1,-\frac{1}{2}}^{\frac{3}{2}} \\ &= C\|v\|_{\frac{1}{2},s}\|w\|_{0,-\frac{1}{2}}^{\frac{1}{2}}\|w\|_{1,-\frac{1}{2}}^{\frac{3}{2}}.\end{aligned}$$

Collecting relations (4), (5), (6), (8), (9), (10), and (14), we get

$$\begin{aligned}&\|w(t)\|_{0,-\frac{1}{2}}^2 + 2\nu\int_0^t \left(\|\partial_1 w\|_{0,-\frac{1}{2}}^2 + \|\partial_2 w\|_{0,-\frac{1}{2}}^2\right)\mathrm{d}\tau \\ &\le C\int_0^t \|w\|_{1,-\frac{1}{2}}^{\frac{3}{2}}\|w\|_{0,-\frac{1}{2}}^{\frac{1}{2}}(\|v\|_{\frac{1}{2},s} + \|\widetilde{v}\|_{\frac{1}{2},s})\,\mathrm{d}\tau \\ &+ C\int_0^t \|w\|_{1,-\frac{1}{2}}^{\frac{1}{2}}\|w\|_{0,-\frac{1}{2}}^{\frac{3}{2}}\|\widetilde{v}\|_{\frac{1}{2},s}\,\mathrm{d}\tau.\end{aligned}$$

Using that $ab \leq \frac{a^4}{4} + \frac{3b^{\frac{4}{3}}}{4}$ for suitable choices of a and b, we infer that

$$\begin{aligned}\|w(t)\|^2_{0,-\frac{1}{2}} + 2\nu \int_0^t \left(\|\partial_1 w\|^2_{0,-\frac{1}{2}} + \|\partial_2 w\|^2_{0,-\frac{1}{2}}\right) \mathrm{d}\tau \leq \nu \int_0^t \|w\|^2_{1,-\frac{1}{2}} \,\mathrm{d}\tau \\ + C \int_0^t \|w\|^2_{0,-\frac{1}{2}} \left(\|v\|^4_{\frac{1}{2},s} + \|\widetilde{v}\|^4_{\frac{1}{2},s} + \|\widetilde{v}\|^{\frac{4}{3}}_{\frac{1}{2},s}\right) \mathrm{d}\tau.\end{aligned}$$

As

$$\|w\|^2_{1,-\frac{1}{2}} = \|\partial_1 w\|^2_{0,-\frac{1}{2}} + \|\partial_2 w\|^2_{0,-\frac{1}{2}} + \|w\|^2_{0,-\frac{1}{2}},$$

we further deduce that

$$\|w(t)\|^2_{0,-\frac{1}{2}} \leq \int_0^t \|w(\tau)\|^2_{0,-\frac{1}{2}} h(\tau) \,\mathrm{d}\tau, \tag{15}$$

where

$$h(t) = \nu + C(\|v\|^4_{\frac{1}{2},s} + \|\widetilde{v}\|^4_{\frac{1}{2},s} + \|\widetilde{v}\|^{\frac{4}{3}}_{\frac{1}{2},s}).$$

By interpolation,

$$\|v\|_{\frac{1}{2},s} \leq \|v\|^{\frac{1}{2}}_{0,s} \|v\|^{\frac{1}{2}}_{1,s}. \tag{16}$$

The hypothesis made on v implies that $\|v\|_{\frac{1}{2},s} \in \mathrm{L}^4(0,T)$. The same holds for $\widetilde{v}$, so $h \in \mathrm{L}^1(0,T)$. Gronwall's lemma applied in (15) now implies that $w \equiv 0$. The proof is completed.

Acknowledgment. The author is grateful to Jean-Yves Chemin for his careful reading of the manuscript and many helpful discussions.

REFERENCES

[1] M. Cannone, *Ondelettes, Paraproduits et Navier-Stokes*, Diderot Éditeur, Paris, 1995.

[2] J.-Y. Chemin, B. Desjardins, I. Gallagher, and E. Grenier, *Fluids with anisotropic viscosity*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 315–335.

[3] H. Fujita and T. Kato, *On the nonstationary Navier-Stokes system*, Rend. Sem. Mat. Univ. Padova, 32 (1962), pp. 243–260.

[4] D. Iftimie, *The 3D Navier-Stokes equations seen as a perturbation of the 2D Navier-Stokes equations*, Bull. Soc. Math. France, 127 (1999), pp. 473–517.

[5] D. Iftimie, *The resolution of the Navier-Stokes equations in anisotropic spaces*, Rev. Mat. Iberoamericana, 15 (1999), pp. 1–36.

[6] H. Koch and D. Tataru, *Well-posedness for the Navier-Stokes equations*, Adv. Math., 157 (2001), pp. 22–35.

[7] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications,* Vol. 1, Dunod, Paris, 1968.

[8] J. Rauch and M. Reed, *Nonlinear microlocal analysis of semilinear hyperbolic systems in one space dimension*, Duke Math. J., 49 (1982), pp. 397–475.

[9] M. Sablé-Tougeron, *Régularité microlocale pour des problèmes aux limites non linéaires*, Ann. Inst. Fourier (Grenoble), 36 (1986), pp. 39–82.